UNIVERSITI PUTRA MALAYSIA

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

NORSIDA BINTI HASAN

IPM 2014 8
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
COPYRIGHT

All material contained within the thesis, including without limitation text, logos, icons, photographs and all other artwork, is copyright material of Universiti Putra Malaysia unless otherwise stated. Use may be made of any material contained within the thesis for non-commercial purposes from the copyright holder. Commercial use of material may only be made with the express, prior, written permission of Universiti Putra Malaysia.

Copyright © Universiti Putra Malaysia
DEDICATIONS

To my beloved husband, Abd Wahab Jusoh,
parents, Hasan Omar and Diwi Che Mat,
sisters, Ruzana and Siti Nur.

Thank you for all of your support along the way.
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 have 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 mechanism; missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Imputation using classification tree outperform the imputation 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.
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


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

In the name of Allah, the Most Gracious and the Most Merciful.

Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. Special appreciation goes to my supervisor, Associate Professor Dr. Mohd Bakri Adam, for his supervision and constant support. This thesis would have not be possible without his guidance and support for the last four years. Not forgotten, my appreciation to my co-supervisors, Associate Professor Dr. Mohd Rizam Abu Bakar and Associate Professor Dr. Norwati Mustapha for their support and knowledge regarding this topic. I would like to express my gratitude to Universiti Pendidikan Sultan Idris for providing data for this research. I met many great people during my four years stay at Universiti Putra Malaysia, their kindness means a lot to me. Thank you very much.
I certify that a Thesis Examination Committee has met on (18 December 2014) to conduct the final examination of (Norsida binti Hasan) on his (or her) thesis entitled “Student’s Background and Academic Performance Modeling with Missing Values using Classification Tree” in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the (Degree of Doctor of Philosophy).

Members of the Thesis Examination Committee were as follows:

Mat Rofa b Ismail, Ph.D.
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairperson)

Noor Akma bt Ibrahim, Ph.D.
Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Md Nasir b Sulaiman, Ph.D.
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Mojtaba Ganjali, Ph.D.
Professor
Faculty of Mathematical Sciences
Shahid Beheshti University
Iran
(External Examiner)

ZULKARNAIN ZAINAL, Ph.D.
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

vi
This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

The members of the Supervisory Committee were as follows:

**Mohd Bakri Adam, Ph.D.**  
Associate Professor  
Institute for Mathematical Research (INSPEM)  
Universiti Putra Malaysia  
(Chairperson)

**Mohd Rizam Abu Bakar, Ph.D.**  
Associate Professor  
Faculty of Science  
Universiti Putra Malaysia  
(Member)

**Norwati Mustapha, Ph.D.**  
Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Member)

---

**BUJANG KIM HUAT, Ph.D.**  
Professor and Dean  
School of Graduate Studies  
Universiti Putra Malaysia  

Date:

vii
DECLARATION

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institution;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification / fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ______________________  Date: ______________________

Name and Matric No.: NORSIDA BINTI HASAN (GS23151)
Declaration by Members of Supervisory Committee

This is to confirm that

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 202-2013) are adhered to

Signature: ___________________________  Signature: ___________________________
Name of Chairman of Supervisory Committee: ___________________________
Name of Member of Supervisory Committee: ___________________________

Signature: ___________________________  Signature: ___________________________
Name of Member of Supervisory Committee: ___________________________

# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>APPROVAL</td>
<td>vi</td>
</tr>
<tr>
<td>DECLARATION</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xiii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xv</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xviii</td>
</tr>
</tbody>
</table>

## CHAPTER

### 1 INTRODUCTION
- 1.1 Student’s Academic Performance .......................... 1
- 1.2 Classification Tree ..................................... 1
- 1.3 Problem Statements ...................................... 2
- 1.4 Research Objectives ...................................... 4
- 1.5 Research Contributions ................................... 4
- 1.6 Organization of Thesis ................................... 4

### 2 LITERATURE REVIEW
- 2.1 Factors Affecting Academic Performance .................. 7
- 2.2 Meta Analysis of Students’ Performance Between Gender .... 8
- 2.3 Predicting Academic Performance Using Classification and Regression Tree .................. 8
- 2.4 Missing Data and Imputation using Classification Tree ...... 10
- 2.5 Conclusion .................................................. 11

### 3 RESEARCH METHODOLOGY
- 3.1 Introduction ................................................ 13
- 3.2 Research Framework ........................................ 13
- 3.3 Data Collection ............................................ 13
- 3.4 Data Pre-processing and Missing Data Injection ............ 16
  - 3.4.1 Data Selection and Transformation ..................... 16
- 3.5 Model Design ................................................ 18
- 3.6 Model Development .......................................... 18
- 3.7 Model Implementation and Evaluation ........................ 18
  - 3.7.1 Cross Validation ....................................... 19
  - 3.7.2 Confusion Matrix ...................................... 19
- 3.8 Conclusion .................................................. 20
4 DATA PRE-PROCESSING AND MISSING DATA INJECTION

4.1 Descriptive Analysis on Students Admission

4.2 Descriptive Analysis on Students Performance

4.2.1 Performance According to Faculty
4.2.2 Performance According to Intake Category
4.2.3 Performance According to Gender
4.2.4 Performance According to Age Group
4.2.5 Performance According to Race
4.2.6 Performance According to Gender and Faculty
4.2.7 Performance According to Gender and Intake Category
4.2.8 Performance According to Age Group and Race
4.2.9 Performance According to Age Group and Gender
4.2.10 Performance According to Age Group and Faculty
4.2.11 Performance According to Age Group and Intake Category
4.2.12 Performance According to Race and Faculty
4.2.13 Performance According to Race and Intake Category

4.3 Data Analysis of Academic Performance Using Meta-Analysis

4.4 Meta-Analysis for First Class Degree Classification
4.5 Meta-Analysis for Second Class Upper Degree Classification
4.6 Meta-Analysis for Second Class Lower Degree Classification

4.7 Mining Students’ Academic Performance using Classification Tree, Bayesian Network and Naive Bayes

4.8 Simulation of Population Data

4.8.1 Algorithm for Simulation of Population Data

4.9 Missing Data Injection

4.9.1 Missing Data Mechanism
4.9.2 Missing Completely at Random (MCAR)
4.9.3 Missing at Random (MAR)
4.9.4 Missing Not at Random (MNAR)

4.10 The influence of Missing Data in Classification Tree, Bayesian network and Naive Bayes

4.11 Sensitivity of Missing Value in Classification Tree using Simulated Dataset

4.12 Conclusion

5 MODEL DEVELOPMENT

5.1 Introduction

5.2 Development of Classification Tree with Imputation Model

5.2.1 Algorithm for Classification Tree with Imputation Model
5.2.2 Algorithm for Missing Values Imputation

5.3 Conclusion

6 EXPERIMENTAL RESULTS

6.1 Introduction

6.2 Result of Imputation using Classification Tree

6.3 Result of Classification Tree with Imputation using Bayesian Network
6.4 Result of Imputation using Naive Bayes 84
6.5 Classification Tree Model with Imputation 86
6.6 Conclusion 87

7 CONCLUSION AND FUTURE WORK 88
7.1 Conclusion 88
7.2 Suggestion for Future Research 89

REFERENCES/BIBLIOGRAPHY 90
APPENDICES 94
BIODATA OF STUDENT 96
LIST OF PUBLICATIONS 97
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Format of Students Data</td>
</tr>
<tr>
<td>3.2</td>
<td>Background of Students</td>
</tr>
<tr>
<td>3.3</td>
<td>Example of a confusion matrix for binary prediction</td>
</tr>
<tr>
<td>4.1</td>
<td>Cross tabulation of faculty and degree classification</td>
</tr>
<tr>
<td>4.2</td>
<td>Cross tabulation of intake category and degree classification</td>
</tr>
<tr>
<td>4.3</td>
<td>Cross tabulation of intake category and degree classification (continue)</td>
</tr>
<tr>
<td>4.4</td>
<td>Cross tabulation of gender and degree classification</td>
</tr>
<tr>
<td>4.5</td>
<td>Cross tabulation of age group and degree classification</td>
</tr>
<tr>
<td>4.6</td>
<td>Cross tabulation of race and degree classification</td>
</tr>
<tr>
<td>4.7</td>
<td>Descriptive statistics of male and female students in eight faculties</td>
</tr>
<tr>
<td>4.8</td>
<td>Meta-analysis for first class degree between female and male students</td>
</tr>
<tr>
<td>4.9</td>
<td>Meta-analysis for second class upper degree between female and male students</td>
</tr>
<tr>
<td>4.10</td>
<td>Meta-analysis for second class lower degree between female and male students</td>
</tr>
<tr>
<td>4.11</td>
<td>Classification Rules for the Left Branch</td>
</tr>
<tr>
<td>4.12</td>
<td>Classification Rules for the students entering the university at the age 29 or below</td>
</tr>
<tr>
<td>4.13</td>
<td>Classification rules for the students entering the university at the age 30 or above</td>
</tr>
<tr>
<td>4.14</td>
<td>Classification accuracy for classification tree, bayesian network and naive bayes</td>
</tr>
<tr>
<td>4.15</td>
<td>Confusion matrix for degree classification using classification tree</td>
</tr>
<tr>
<td>4.16</td>
<td>Class wise accuracy for three classes prediction using classification tree</td>
</tr>
<tr>
<td>4.17</td>
<td>Confusion matrix for degree classification using Bayesian network</td>
</tr>
<tr>
<td>4.18</td>
<td>Class wise accuracy for three classes prediction using Bayesian network</td>
</tr>
<tr>
<td>4.19</td>
<td>Confusion matrix for degree classification using naive Bayes</td>
</tr>
<tr>
<td>4.20</td>
<td>Class wise accuracy for three classes prediction using naive Bayes</td>
</tr>
<tr>
<td>4.21</td>
<td>Summary of simulation data</td>
</tr>
<tr>
<td>4.22</td>
<td>The 95% confidence interval of accuracy for classification tree, bayesian network and naive Bayes</td>
</tr>
<tr>
<td>4.23</td>
<td>Summary of the tree models at different level of missing values when missing values occur in variable FACULTY</td>
</tr>
<tr>
<td>4.24</td>
<td>Summary of the tree models at different level of missing values when missing values occur in variables FACULTY and CATEGORY</td>
</tr>
<tr>
<td>4.25</td>
<td>Summary of the tree models at different level of missing values when missing values occur in variables FACULTY, CATEGORY and AGE GROUP</td>
</tr>
</tbody>
</table>

xiii
4.26 Summary of the tree models at different level of missing values when missing values occur in variables FACULTY, CATEGORY, AGE GROUP and RACE

4.27 The 95% confidence interval of the tree models with different level of MCAR

4.28 The 95% confidence interval of the tree models for MAR and MNAR

6.1 Summary of the classification tree model before and after imputation for MCAR using classification tree

6.2 Summary of the classification tree model before and after imputation for MAR

6.3 Summary of the classification tree model before and after imputation for MNAR

6.4 Summary of the classification tree model before and after imputation for MCAR using Bayesian network

6.5 Summary of the classification tree model before and after imputation for MAR using Bayesian network

6.6 Summary of the classification tree model before and after imputation for MNAR using Bayesian network

6.7 Summary of the classification tree model before and after imputation for MCAR using naive Bayes

6.8 Summary of the classification tree model before and after imputation for MAR using naive Bayes

6.9 Summary of the classification tree model before and after imputation for MNAR using naive Bayes
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Research Framework</td>
<td>14</td>
</tr>
<tr>
<td>4.1 Pie chart of students admission according to faculty</td>
<td>21</td>
</tr>
<tr>
<td>4.2 Pie chart of students admission according to gender</td>
<td>22</td>
</tr>
<tr>
<td>4.3 Pie chart of students admission according to age group</td>
<td>22</td>
</tr>
<tr>
<td>4.4 Mosaic plot of students admission according to faculty and age group</td>
<td>23</td>
</tr>
<tr>
<td>4.5 Bar chart of students admission according to state</td>
<td>23</td>
</tr>
<tr>
<td>4.6 Bar chart of students admission according to intake category</td>
<td>24</td>
</tr>
<tr>
<td>4.7 Pie chart of students degree classification</td>
<td>25</td>
</tr>
<tr>
<td>4.8 Degree classification by gender</td>
<td>31</td>
</tr>
<tr>
<td>4.9 Mosaic plot of degree classification by gender and faculty</td>
<td>33</td>
</tr>
<tr>
<td>4.10 Mosaic plot of degree classification by gender and intake category</td>
<td>34</td>
</tr>
<tr>
<td>4.11 Mosaic plot of degree classification by age group and race</td>
<td>35</td>
</tr>
<tr>
<td>4.12 Mosaic plot of degree classification by age group and gender</td>
<td>36</td>
</tr>
<tr>
<td>4.13 Mosaic plot of degree classification by age group and faculty</td>
<td>37</td>
</tr>
<tr>
<td>4.14 Mosaic plot of degree classification by age group and intake category</td>
<td>38</td>
</tr>
<tr>
<td>4.15 Mosaic plot of degree classification by race group and faculty</td>
<td>39</td>
</tr>
<tr>
<td>4.16 Mosaic plot of degree classification by race and faculty</td>
<td>39</td>
</tr>
<tr>
<td>4.17 Mosaic plot of degree classification according to race group and intake category</td>
<td>40</td>
</tr>
<tr>
<td>4.18 Mosaic plot of degree classification by race and intake category</td>
<td>41</td>
</tr>
</tbody>
</table>
4.19 Forest plot of observed effect sizes and the 95% confidence intervals for the first class degree classification studies. 45
4.20 Funnel plots for the first class degree classification studies. 46
4.21 Forest plot of observed effect sizes and the 95% confidence intervals for the second class upper degree classification studies. 48
4.22 Funnel plots for the second class upper degree classification studies. 48
4.23 Forest plot of observed effect sizes and the 95% confidence intervals for the second class lower degree classification studies. 50
4.24 Funnel plots for the second class lower degree classification studies. 51
4.25 Classification tree model of students degree classification 53
4.26 Naive Bayes classification model of students degree classification 57
4.27 Bayesian network classification model of students degree classification 58
4.28 Mosaic plot of students Degree Classification using real dataset 64
4.29 Mosaic plot of students Degree Classification using simulation dataset 64
4.30 Classification tree model of students degree classification using simulation data 65
4.31 Percentage of correct classification rate for classification tree, bayesian network and naive Bayes with different level of missing values 67
4.32 Percentage of correct classification rate in dataset with different level of missing values 70
4.33 Percentage of correct classification rate in dataset with different level of missing values 71
5.1 Classification tree used to impute the missing value in variable FACULTY 74
5.2 Classification Tree to impute missing data in variable FACULTY 75
5.3 Classification Tree to impute missing data in variable AGE GROUP 76
5.4 Classification Tree to impute missing data in variable RACE 76
5.5 Classification Tree to impute missing data in variable CATEGORY 77
5.6 Bayesian network learnt from complete sub-dataset 78
5.7 Naive Bayes classifier to impute missing data in variable FACTULTY

6.1 Classification tree model with imputation for students degree classification

6.2 Comparison of classification accuracy after imputation using classification tree, Bayesian network and naive Bayes
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>STPM</td>
<td>Malaysian Higher School Certificate</td>
</tr>
<tr>
<td>PKPG</td>
<td>In-service Teacher Education Programme</td>
</tr>
<tr>
<td>KDPK</td>
<td>In-service Teachers with Diploma in Special Education Programme</td>
</tr>
<tr>
<td>MCAR</td>
<td>Missing Completely At Random</td>
</tr>
<tr>
<td>MAR</td>
<td>Missing At Random</td>
</tr>
<tr>
<td>MNAR</td>
<td>Missing Not At Random</td>
</tr>
<tr>
<td>RRP</td>
<td>Random Recursive Partitioning</td>
</tr>
<tr>
<td>ITree</td>
<td>Imputation Tree</td>
</tr>
<tr>
<td>UPSI</td>
<td>Universiti Pendidikan Sultan Idris</td>
</tr>
<tr>
<td>FB</td>
<td>Faculty of Languages</td>
</tr>
<tr>
<td>FPE</td>
<td>Faculty of Business and Economics</td>
</tr>
<tr>
<td>FSKPM</td>
<td>Faculty of Cognitive Science and Human Development</td>
</tr>
<tr>
<td>FSM</td>
<td>Faculty of Music</td>
</tr>
<tr>
<td>FSS</td>
<td>Faculty of Sports Science</td>
</tr>
<tr>
<td>FSSK</td>
<td>Faculty of Human Sciences</td>
</tr>
<tr>
<td>FST</td>
<td>Faculty of Science and Technology</td>
</tr>
<tr>
<td>FTMK</td>
<td>Faculty of Information Technology and Communication</td>
</tr>
<tr>
<td>CGPA</td>
<td>Cumulative Grade Point Average</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
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 meaningful 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 $\Delta i_t$ 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

$$P_l i(t_l) + P_r i(t_r).$$

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 classification 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 mechanism; 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 contribution 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 research contribution.

Chapter 2 presents 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 explanation 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.


91


Ssali, G. and Marwala, T. 2008. Computational Intelligence and Decision Tree for Missing Data Estimation. In IEEE World Congress on Computational Intelligence Proceeding, 201–207.


