05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi



UNIVERSITI PUTRA MALAYSIA

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

December 2014

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi
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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

UPD DONE MyTO SCAN

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi
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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

TABLE OF CONTENTS

	Page
ABSTRACT	i
ABSTRAK	iii
ACKNOWLEDGEMENTS	v
APPROVAL	vi
DECLARATION	viii
LIST OF TABLES	xiii
LIST OF FIGURES	xv
LIST OF ABBREVIATIONS	xviii
CHAPTER	
1 **INTRODUCTION**	1
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**	7
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**	13
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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4 DATA PRE-PROCESSING AND MISSING DATA INJECTION 21
4.1 Descriptive Analysis on Students Admission 21
4.2 Descriptive Analysis on Students Performance 24
4.2.1 Performance According to Faculty 25
4.2.2 Performance According to Intake Category 27
4.2.3 Performance According to Gender 30
4.2.4 Performance According to Age Group 31
4.2.5 Performance According to Race 32
4.2.6 Performance According to Gender and Faculty 33
4.2.7 Performance According to Gender and Intake Category 33
4.2.8 Performance According to Age Group and Race 35
4.2.9 Performance According to Age Group and Gender 36
4.2.10 Performance According to Age Group and Faculty 37
4.2.11 Performance According to Age Group and Intake Category 38
4.2.12 Performance According to Race and Faculty 38
4.2.13 Performance According to Race and Intake Category 40
4.3 Data Analysis of Academic Performance Using Meta-Analysis 42
4.4 Meta-Analysis for First Class Degree Classification 44
4.5 Meta-Analysis for Second Class Upper Degree Classification 47
4.6 Meta-Analysis for Second Class Lower Degree Classification 49
4.7 Mining Students' Academic Performance using Classification Tree, Bayesian Network and Naive Bayes 52
4.8 Simulation of Population Data 61
4.8.1 Algorithm for Simulation of Population Data 62
4.9 Missing Data Injection 64
4.9.1 Missing Data Mechanism 65
4.9.2 Missing Completely at Random (MCAR) 65
4.9.3 Missing at Random (MAR) 66
4.9.4 Missing Not at Random (MNAR) 66
4.10 The influence of Missing Data in Classification Tree, Bayesian network and Naive Bayes 67
4.11 Sensitivity of Missing Value in Classification Tree using Simulated Dataset 68
4.12 Conclusion 72

5 MODEL DEVELOPMENT 73
5.1 Introduction 73
5.2 Development of Classification Tree with Imputation Model 73
5.2.1 Algorithm for Classification Tree with Imputation Model 73
5.2.2 Algorithm for Missing Values Imputation 74
5.3 Conclusion 80

6 EXPERIMENTAL RESULTS 81
6.1 Introduction 81
6.2 Result of Imputation using Classification Tree 81
6.3 Result of Classification Tree with Imputation using Bayesian Network 83

05-4506832 Perpustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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 89
7.1 Conclusion 89
7.2 Suggestion for Future Research 90

REFERENCES/BIBLIOGRAPHY 91
APPENDICES 95
BIODATA OF STUDENT 97
LIST OF PUBLICATIONS 98

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

LIST OF TABLES

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table		Page
3.1	Format of Students Data	15
3.2	Background of Students	17
3.3	Example of a confusion matrix for binary prediction	19
4.1	Cross tabulation of faculty and degree classification	26
4.2	Cross tabulation of intake category and degree classification	28
4.3	Cross tabulation of intake category and degree classification (continue)	29
4.4	Cross tabulation of gender and degree classification	30
4.5	Cross tabulation of age group and degree classification	31
4.6	Cross tabulation of race and degree classification	32
4.7	Descriptive statistics of male and female students in eight faculties	43
4.8	Meta-analysis for first class degree between female and male students	44
4.9	Meta-analysis for second class upper degree between female and male students	47
4.10	Meta-analysis for second class lower degree between female and male students	49
4.11	Classification Rules for the Left Branch	54
4.12	Classification Rules for the students entering the university at the age 29 or below	55
4.13	Classification rules for the students entering the university at the age 30 or above	56
4.14	Classification accuracy for classification tree, bayesian network and naive bayes	59
4.15	Confusion matrix for degree classification using classification tree	59
4.16	Class wise accuracy for three classes prediction using classification tree	59
4.17	Confusion matrix for degree classification using Bayesian network	60
4.18	Class wise accuracy for three classes prediction using Bayesian network	60
4.19	Confusion matrix for degree classification using naive Bayes	60
4.20	Class wise accuracy for three classes prediction using naive Bayes	61
4.21	Summary of simulation data	63
4.22	The 95% confidence interval of accuracy for classification tree, bayesian network and naive Bayes	67
4.23	Summary of the tree models at different level of missing values when missing values occur in variable FACULTY	68
4.24	Summary of the tree models at different level of missing values when missing values occur in variables FACULTY and CATEGORY	68
4.25	Summary of the tree models at different level of missing values when missing values occur in variables FACULTY, CATEGORY and AGE GROUP	68

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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 68
4.27 The 95% confidence interval of the tree models with different level of MCAR 69
4.28 The 95% confidence interval of the tree models for MAR and MNAR 69

6.1 Summary of the classification tree model before and after imputation for MCAR using classification tree 81
6.2 Summary of the classification tree model before and after imputation for MAR 82
6.3 Summary of the classification tree model before and after imputation for MNAR 82
6.4 Summary of the classification tree model before and after imputation for MCAR using Bayesian network 83
6.5 Summary of the classification tree model before and after imputation for MAR using Bayesian network 83
6.6 Summary of the classification tree model before and after imputation for MNAR using Bayesian network 84
6.7 Summary of the classification tree model before and after imputation for MCAR using naive Bayes 84
6.8 Summary of the classification tree model before and after imputation for MAR using naive Bayes 85
6.9 Summary of the classification tree model before and after imputation for MNAR using naive Bayes 85

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

LIST OF FIGURES

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Figure		Page
3.1	Research Framework	14
4.1	Pie chart of students admission according to faculty	21
4.2	Pie chart of students admission according to gender	22
4.3	Pie chart of students admission according to age group	22
4.4	Mosaic plot of students admission according to faculty and age group	23
4.5	Bar chart of students admission according to state	23
4.6	Bar chart of students admission according to intake category	24
4.7	Pie chart of students degree classification	25
4.8	Degree classification by gender	31
4.9	Mosaic plot of degree classification by gender and faculty	33
4.10	Mosaic plot of degree classification by gender and intake category	34
4.11	Mosaic plot of degree classification by age group and race	35
4.12	Mosaic plot of degree classification by age group and gender	36
4.13	Mosaic plot of degree classification by age group and faculty	37
4.14	Mosaic plot of degree classification by age group and intake category	38
4.15	Mosaic plot of degree classification by race group and faculty	39
4.16	Mosaic plot of degree classification by race and faculty	39
4.17	Mosaic plot of degree classification according to race group and intake category	40
4.18	Mosaic plot of degree classification by race and intake category	41

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

5.7 Naive Bayes classifier to impute missing data in variable FACULTY 79

6.1 Classification tree model with imputation for students degree classification 86

6.2 Comparison of classification accuracy after imputation using classification tree, Bayesian network and naive Bayes 87

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi
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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi
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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun PustakaTBainun ptbupsi

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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun PustakaTBainun ptbupsi

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).

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

CHAPTER 3

RESEARCH METHODOLOGY

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

3.1 Introduction

This chapter describe the mothodology applied in this study. Section 3.2 present the research framework, followed by the data collection in Section 3.3 and data processing in Section 3.4. The model design, model development and model implementation are briefly explained in Sections 3.5, 5.2 and 3.7.

3.2 Research Framework

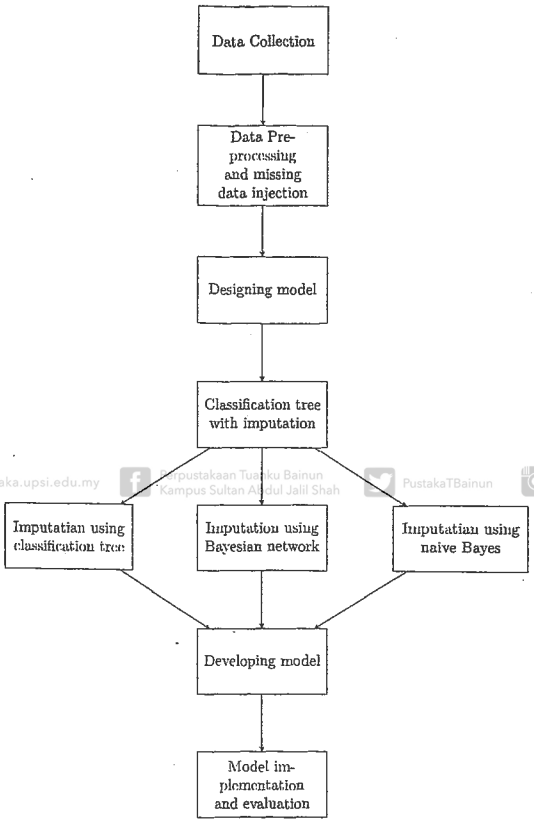
The method used in this study are adopted from previous research in the same field with some modifications. Figure 3.1 outlines the research framework employed in this study which are elaborated in the next section.

3.3 Data Collection

In this study we use undergraduate students data obtained from Universiti Pendidikan Sultan Idris (UPSI) who's graduated from 1997 to 2008. The data is already in a flat file and consist of 22,642 students from eight faculties namely; Faculty of Languages (FB), Faculty of Business and Economics (FPE), Faculty of Cognitive Science and Human Development (FSKPM), Faculty of Music (FSM), Faculty of Sports Science (FSS), Faculty of Human Sciences (FSSK), Faculty of Science and Technology (FST) and Faculty of Information Technology and Communication (FTMK). The usage of large sample size in the data mining analysis will increase the accuracy of prediction (Vialardi et al., 2009). The dataset is in format of the original data is shown in Table 3.1.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi



1

Figure 3.1: Research Framework

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 3.1: Format of Students Data

No	Field Name	Field Type	No. of Categories	List of Categories
1	Identification Number	Categorical	-	-
2	Name	Categorical	-	-
3	Gender	Categorical	2	Male, Female
4	Race	Categorical	8	Malay, Chinese, Indian, Foreigner, Origin Native, Bumiputra Sabah, Bumiputra Sarawak, Others
5	State	Categorical	17	Johor, Kedah, Kelantan, Overseas, Melaka, Negeri Sembilan, Pahang, Perak, Perlis, Pulau Pinang, Sabah, Sarawak, Selangor, Terengganu, Kuala Lumpur, Labuan, Putrajaya.
6	Intake Channel	Categorical	16	Institut Aminuddin Baki, Special Case/Diploma (Science), Special Case/Diploma (Arts), Matriculation (Accounts), Matriculation (Science), KDPK, PKPG (College), PKPG (UPSI), Overseas, Teacher Certificate, MPSAH, STPM (Science), STPM (Arts), Second Channel, Change University
7	Faculty	Categorical	8	FB, FPE, FSKPM, FSM, FSS, FSSK, FST, FTMK.
8	Age	Continuous	-	23 - 60 Tahun
9	CGPA	Continuous	-	2.00 - 3.92

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

3.4 Data Pre-processing and Missing Data Injection

The data ontained from UPSI is already in a flat file and contain no missing values. Firstly, select the variables that will be use in the model namely gender, race, state, intake channel, faculty age and CGPA. Secondly, we transform all the numerical data values to categorical data values. The data selection and transformation is further describe is Section 3.4.1

3.4.1 Data Selection and Transformation

The students' background information or the predictors are categorical data consist of identification number, name, gender, age, faculty, intake category, race, religion and state while the response variable is the Cumulative Gred Point Average (CGPA). Students' CGPA are then converted to the degree classification. Students are classified either in first class (A), second class upper division (B) or second class lower division (C) according to the university degree classification scale. The students' name and their identification numbers are not use in the analysis and classification tree model. We also convert the student age into four categories. Students who are enter the institution at or below age 29 are in group 1 (U1), age 30 - 39 are in group 2 (U2), age 40 - 49 are in group 3 (U3) and age 50 and above are in group 4 (U4). The summary of the students' background data is shown in Table 4.21. The data is further analyzed in Chapter 4.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 3.2: Background of Students

Variable		Percentage
Gender	Male	34.39
	Female	65.61
Race	Malay	85.02
	Chinese	5.67
	Indian	3.86
	Bumiputra Sabah	2.69
	Bumiputra Sarawak	2.06
	Foreigner	0.13
	Native Origin	0.01
	Others	0.56
Age Group	29 or below (U1)	50.23
	30 - 39 (U2)	27.49
	40 - 49 (U3)	18.28
	50 and above (U4)	4.00
Faculty	Faculty of Languages	12.86
	Faculty of Management and Economics	12.90
	Faculty of Cognitive Science and Human Development	15.32
	Faculty of Music	9.23
	Faculty of Sports Science	7.61
	Faculty of Human Sciences	2.88
	Faculty of Science and Technology	29.15
	Faculty of Information Technology and Communication	10.05
Intake Category	Diploma	9.67
	Special Case/Diploma	12.27
	Matriculation	18.60
	STPM	26.25
	Institut Aminudin Baki	3.45
	PKPG	16.76
	KDPK	4.25
	Second Channel	7.70
	Others	1.05
Degree Classification	First Class (A)	0.65
	Second Class Upper Division (B)	59.95
	Second Class Lower Division (C)	39.40

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

3.5 Model Design

In this research, missing values is imputed using three imputation methods namely classification tree, Bayesian network and naive Bayes. Each types of missing data mechanism; MCAR, MAR and MNAR is imputed using classification tree, Bayesian network and naive Bayes. After that, the method that produce the most accurate imputation is chosen to model the classification tree with imputation.

Naive Bayes method is a simple probabilistic classifier that apply Bayes theorem with the strong (naive) assumption of independence between each pair of features (Han and Kamber, 2001; Bekele and Menzel, 2005). The naive Bayes classifier combines Bayes theorem with a decision rule known as maximum a posteriori or MAP decision rule. Assume that y is a class variable and $x_1, ..., x_n$ is an independent vector, Bayes theorem can be written as

$$P(y|x_1, ..., x_n) = \frac{P(y)P(x_1,...,x_n|y)}{P(x_1,...,x_n)}$$

The naive Bayes classifier is then can be written as

$$\widehat{y} = arg_{max} P(y) \prod p(x_i, ..., x_n|y)$$

Bayesian network is a probabilistic graphical model that involves the theory of how to minimize risk and loss when making decisions based on uncertain information (Han and Kamber, 2001). The structure of a Bayesian network illustrates the interactions among the set of variables that it models. Bayesian network model consists of a directed acyclic graph that represents a set of random variables and conditional probability distributions.

3.6 Model Development

After choosing the best imputation model, Classification tree with imputation model proposed in 3.5 is developed using R language. The model development is further discussed in Chapter 5.

3.7 Model Implementation and Evaluation

Classification tree with imputation model is implement to model the real data of student's background and academic performance.
The result is then evaluated to obtain the performance of the classification tree with imputation. We used the percentage of correct classification to evaluate the classification accuracy of the tree model as shown by

$$\text{Classification Accuracy} = \frac{\sum_{i=1}^{m} c_i}{n} \times 100\%. \tag{3.1}$$

c_i is the number of correct classification at leaf node i, m is the total number of leaf nodes and n is the number of all observation. First, we calculate the correct

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

classification rate for all leaf node i, c_i, then, we calculate the percentage of classification accuracy as shown in eq.3.1.

3.7.1 Cross Validation

The 10-fold cross validation is chosen to estimate the true prediction accuracy for this dataset using classification tree. The 10-fold cross validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set.
In this 10-fold cross-validation, the original sample is randomly partitioned into 10 subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The k results from the folds is then averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. The true prediction accuracy is estimated as the average accuracy on the test sample as given by

$$\text{Average Accuracy} = \frac{1}{k}\sum_{i=1}^{k} N_i,$$

where k is the number of the subsample and N_i is the accuracy of the validation sample in each fold.

3.7.2 Confusion Matrix

A confusion matrix is used to display the number of correct and incorrect predictions made by the model compared with the actual class. The matrix is $K \times K$, where K is the number of classes where KEach column of the matrix represents the records in a predicted class, while each row represents the records in an actual class. In this study, we have 3×3 matrix to rpresent 3 predicted classes and 3 actual classes a It is also known as the contingency table or the error matrix. Confusion matrix allows more detailed analysis than mere proportion of correct accuracy because it reports the number of false positive (FP), false negative (FN), true positive (TP), and true negative (TN). The example of confusion matrix for binary prediction is shown in Table 3.3

Table 3.3: Example of a confusion matrix for binary prediction

		Predicted Class	
		Class A	Not Class A
Actual Class	Class A	a	b
	Not Class A	c	d

The true positive (TP) is the proportion of class A cases that are correctly classified as class A and is calculated as

$$TP = \frac{a}{a+b} .$$

The false positive (FP) is the proportion of Not Class A cases that are incorrectly classified as Class A and is calculated as

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

$$FP = \frac{c}{c+d} .$$

The true negative (TN) is defined as the proportion of Not Class A cases that are classified correctly and is calculated as

$$TN = \frac{d}{c+d} .$$

The false negative (FN) is the proportion of Class A cases that were incorrectly classified as Not Class A and is calculated as

$$FN = \frac{b}{a+b} .$$

The accuracy determined using eq.3.1 may not be an adequate performance measure when the gap between the number of Class A cases and Not Class A cases is large (Kubat and Matwin, 1997). A Confusion Matrix on the other hand, is able to indicate the model behaviour for individual classes.

3.8 Conclusion

This chapter explains the methodology used in this research which include the research framework, the requirement analysis, data pre-processing, model design, model development and model implementation.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

CHAPTER 4
DATA PRE-PROCESSING AND MISSING DATA INJECTION

The descriptive data analysis is carried out to investigate the nature of the dataset. Cross tabulation analysis and mosaic plots are conducted to determine the relationship between categorical variables of students' academic performance according to their gender, university academic intake category, age and race. We present a meta-analysis of eight studies of gender differences in academic performance. The differences of gender performance on three level of performance (first class, second class upper and second class lower) are examined. Section 4.7 present the result of mining students'academic performance using classification tree, bayesian network and naive Bayes. Finally we describe the simulation of population data to investigate the effect of missing values in classification tree, bayesian network and naive Bayes in Section 4.8.

4.1 Descriptive Analysis on Students Admission

Figure 4.1 displays the distribution of students according to faculty. FST is the largest faculty in the university with 29.1 percent of students registered followed by FSKPM with 15.3 percent. Both FB and FPE contribute for 12.9 percent of students population. The percentage of other faculties are not more than 10 percent.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

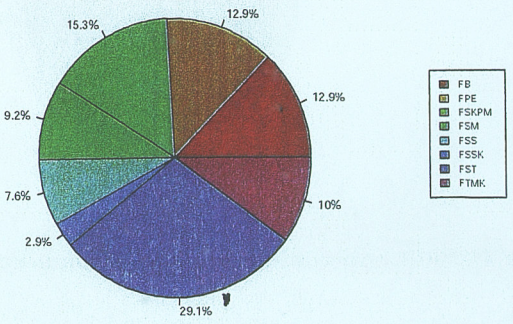


Figure 4.1: Pie chart of students admission according to faculty

Based on the results shown in Figure 4.2 majority of the registered students are females with ratio of 65.6 percent in this university population and almost 34.4 percent are males. The females enrollment rate is more than half of the males enrollment rate.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

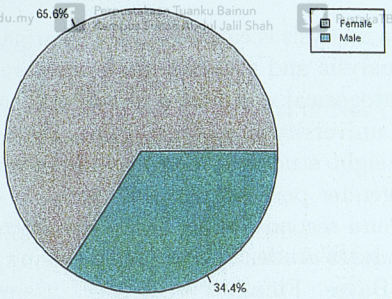


Figure 4.2: Pie chart of students admission according to gender

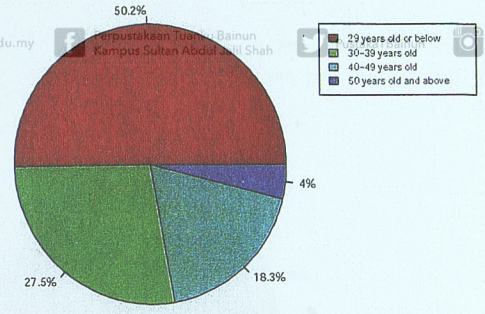


Figure 4.3: Pie chart of students admission according to age group

Figure 4.3 shows that majority of the registered students entering the university at the age 29 years old or below. The percentage of students entering the university decreases for older students.

Figure 4.4 shows the pattern of registered students according to age group in each faculty. Younger students (age 29 years old or below) in general are the majority in all faculties. FSSK, FSKPM and FST have the highest percentage of young students with 70.4 percent, 66.9 percent and 58.1 percent respectively. Students aged 50 years old and above are most likely to register at PFE with 13.4 percent.

FSSK on the other hand has no students with age 50 years old and above.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

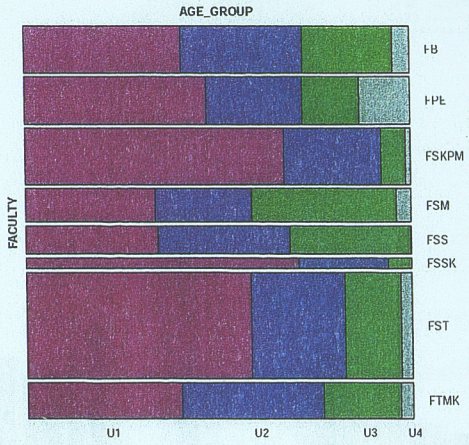


Figure 4.4: Mosaic plot of students admission according to faculty and age group

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

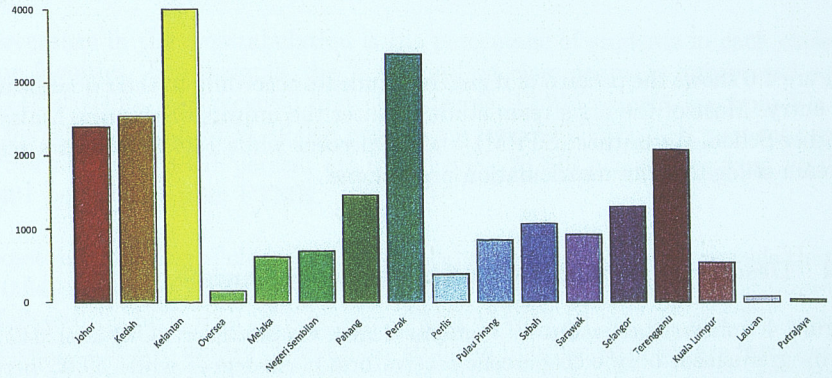
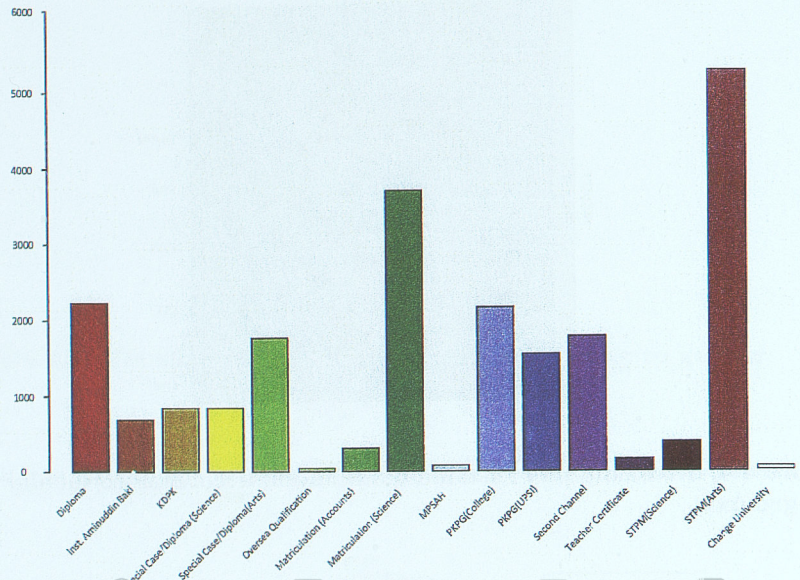


Figure 4.5: Bar chart of students admission according to state

Figure 4.5 shows the pattern of registered students according to their state. Majority of students come from Kelantan with 17.9 percent, followed by Perak with 14.5 percent, Kedah with 11.2 percent, Johor with 10.5 percent and Terengganu with 9.2 percent. Students from Wilayah Persekutuan Putrajaya and Labuan are

the minority in this university.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi



05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Figure 4.6: Bar chart of students admission according to intake category

Figure 4.6 shows the pattern of registered students according to their qualifications at entry. Most of the art stream students enter the university through Malaysian Higher School Certificate (STPM) with 24 percent while 16.8 percent are science stream students from matriculation programme.

4.2 Descriptive Analysis on Students Performance

Figure 4.7 shows the pattern of students degree classification. Out of 22642 students graduated, only 0.65 percent receive first class degree while 59.95 percent receive second class upper division and the remaining 39.40 percent receive second class lower division.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

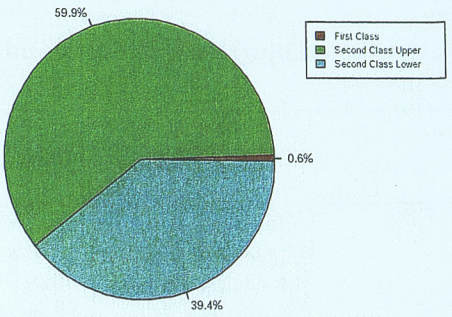


Figure 4.7: Pie chart of students degree classification

4.2.1 Performance According to Faculty

Table 4.1 shows an association between faculty and degree classification. A crosstabulation is a joint frequency distribution of cases based on two or more categorical variables. The % within Faculty in the crosstabulation is the percentage of students in each faculties who received first class degree, second class upper degree and second class lower degree.

The percentage in the crosstabulation is the percentage of students in each class based on faculties. For example, the % within first class degree is interpreted as out of 147 students who received first class degree, 12.246 percent are from FB, 34.014 percent are from FPE, 19.048 percent are from FSKPM, 2.041 percent are from FSM and FSS, 3.401 percent are from FSSK, 23.810 percent are from FST and 3.401 percent are from FTMK.

The crosstabulated degree classification by faculty shown in Table 4.1 reveals that out of 0.649 percent of students who receive first class degree 34 percent are from FPE, followed by FST with 23.8 percent, FSKPM with 19 percent and FB with 12.2 percent. The percentage of first class students in other faculties are less than 4 percent. Comparison among faculties reveal that majority of students in FTMK receive second class lower degree. Out of all 2275 students in FTMK, nearly 60 percent receive second class lower degree whereas out of 653 students in FSSK, only 9.8 percent receive second class lower degree.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.1: Cross tabulation of faculty and degree classification

			Class			
			First	2nd Upper	2nd Lower	Row Total
Faculty	FB	Count	18	1693	1200	2911
		Expected Count	18.9	1745.0	1147.1	
		% within Faculty	0.618	58.159	41.223	
		% within Class	12.245	12.473	13.450	
	FPE	Count	50	2218	652	2920
		Expected Count	19.0	1750.4	1150.6	
		% within Faculty	1.712	75.959	22.329	
		% within Class	34.014	16.341	7.308	
	FSKPM	Count	28	2871	571	3470
		Expected Count	22.5	2080.1	1367.3	
		% within Faculty	0.807	82.738	16.455	
		% within Class	19.048	21.152	6.400	
	FSM	Count	3	1137	949	2089
		Expected Count	13.6	1252.3	823.2	
		% within Faculty	0.144	54.428	45.428	
		% within Class	2.041	8.377	10.637	
	FSS	Count	3	1042	679	1724
		Expected Count	11.2	1033.5	679.3	
		% within Faculty	0.174	60.441	39.385	
		% within Class	2.041	7.677	7.610	
	FSSK	Count	5	584	64	653
		Expected Count	4.2	391.4	257.3	
		% within Faculty	0.766	89.433	9.801	
		% within Class	3.401	4.303	0.717	
	FST	Count	35	3124	3441	6600
		Expected Count	42.8	3956.4	2600.7	
		% within Faculty	0.530	47.333	52.136	
		% within Class	23.810	23.016	38.568	
	FTMK	Count	5	904	1366	2275
		Expected Count	14.8	1363.8	896.5	
		% within Faculty	0.220	39.736	60.044	
		% within Class	3.401	6.660	15.310	
Column Total		Count	147	13573	8922	22642
		% within Class	0.649	59.946	39.405	100

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.2 Performance According to Intake Category

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Tables 4.2 and 4.3 show an association between intake category and degree classification. The crosstabulated degree classification by intake category shown in Tables ?? and 4.3 reveal that out of 0.649 percent of students who receive first class degree 21.8 percent are from Institute of Aminuddin Baki, followed by STPM (Science) with 19.7 percent, Special case/diploma with 12 percent and PKPG (College) with 11.5 percent. The percentage of first class students in other faculties are less than 7 percent. Result also shows that none of the students from KDPK, overseas qualification, teacher certificate and change university receives first class degree.

Comparison among intake categories reveals that majority of students from KDPK and oversea qualification receive second class lower degree. Out of all 962 students from KDPK, almost 75 percent receive second class lower degree and out of 14 students with oversea qualification, almost 71 percent receive second class lower degree. Almost 64 percent of students from both PKPG (UPSI) and second channel receive second class lower degree.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.2: Cross tabulation of intake category and degree classification

			Class			
			First	2nd Upper	2nd Lower	Row Total
Intake Category	Diploma	Count	7	1562	621	2190
		Expected Count	14.2	1312.8	863.0	
		% within Category	0.320	71.324	28.356	
		% within Class	4.762	11.508	6.960	
	Institute of Amiduddin Baki	Count	32	706	44	782
		Expected Count	5.1	468.8	308.1	
		% within Category	4.092	90.281	5.627	
		% within Class	21.769	5.202	0.493	
	KDPK	Count	0	244	718	962
		Expected Count	6.2	576.7	379.1	
		% within Category	0.000	25.364	74.636	
		% within Class	0.000	1.798	8.048	
	Special Case or Diploma (Science)	Count	9	698	272	979
		Expected Count	6.4	586.9	385.8	
		% within Category	0.919	71.297	27.783	
		% within Class	6.122	5.143	3.049	
	Special Case or Diploma (Art)	Count	18	1134	648	1800
		Expected Count	11.7	1079.0	709.3	
		% within Category	1.000	63.000	36.000	
		% within Class	12.245	8.355	7.263	
	Overseas qualification	Count	0	4	10	14
		Expected Count	0.1	8.4	5.5	
		% within Category	0.000	28.571	71.429	
		% within Class	0.000	0.029	0.112	
	Matriculation (Account)	Count	8	305	91	404
		Expected Count	2.6	242.2	159.2	
		% within Category	1.980	75.495	22.525	
		% within Class	5.442	2.247	1.020	
	Matriculation (Science)	Count	6	2130	1671	3807
		Expected Count	24.7	2282.1	1500.1	
		% within Category	0.158	55.950	43.893	
		% within Class	4.082	15.693	18.729	
	MPSAH	Count	5	42	1	48
		Expected Count	0.3	28.8	18.9	
		% within Category	10.417	87.500	2.083	
		% within Class	3.401	0.309	0.011	

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.3: Cross tabulation of intake category and degree classification (continue)

			Class			
			First	2nd Upper	2nd Lower	Row Total
Intake Category	PKPG (college)	Count	17	1237	982	2236
		Expected Count	14.5	1340.4	881.1	
		% within Category	0.760	55.322	43.918	
		% within Class	11.565	9.114	11.007	
	PKPG (UPSI)	Count	4	548	1006	1558
		Expected Count	10.1	934.0	613.9	
		% within Category	0.257	35.173	64.570	
		% within Class	2.721	4.037	11.275	
	Second channel	Count	2	619	1122	1743
		Expected Count	11.3	1044.9	686.8	
		% within Category	0.115	35.513	64.372	
		% within Class	1.361	4.561	12.576	
	Teacher certificate	Count	0	154	20	174
		Expected Count	1.1	104.3	68.6	
		% within Category	0.000	88.506	11.494	
		% within Class	0.000	1.135	0.224	
	STPM (Science)	Count	29	264	222	515
		Expected Count	3.3	308.7	202.9	
		% within Category	5.631	51.262	43.107	
		% within Class	19.728	1.945	2.488	
	STPM (Art)	Count	10	3925	1493	5428
		Expected Count	35.2	3253.9	2138.9	
		% within Category	0.184	72.310	27.506	
		% within Class	6.803	28.918	16.734	
	Change university	Count	0	1	1	2
		Expected Count	0.0	1.2	0.8	
		% within Category	0.000	50.000	50.000	
		% within Class	0.000	0.007	0.011	
Column Total		Count	147	13573	8922	22642
		% within Class	0.649	59.946	39.405	100

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.3 Performance According to Gender

Table 4.4 shows an association between gender and degree classification. Examination of the cell frequencies shows that more than 67.5 percent of the female students receive first class or second class upper degree. The percentage of female students in second class lower division is almost 57 percent.

Table 4.4: Cross tabulation of gender and degree classification

			Class			
			First	2nd Upper	2nd Lower	Row Total
Gender	Female	Count	125	9644	5087	14856
		% within Gender	0.841	64.917	34.242	
		% within Class	85.034	71.053	57.016	
	Male	Count	22	3929	3835	7786
		% within Gender	0.283	50.462	49.225	
		% within Class	14.966	28.947	42.984	
Column Total		Count	147	13573	8922	22642
		% within Class	0.649	59.946	39.405	100

Figure 4.8 provides the distribution of degree performance among students according to their gender. The mosaic plot shows that females are the majority group and they also perform better than males. Out of 0.65 percent of the total students received first class degree, 85 percent are females and only 15 percent are males. The similar trend is also shown for the second class upper division where out of 59.9 percent of the total students, 71 percent are females while the remaining 29 percent are males. The performance gap between males and females is smaller for the second class lower division where out of 39.4 percent of the total students, 57 percent are females and 43 percent are males.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

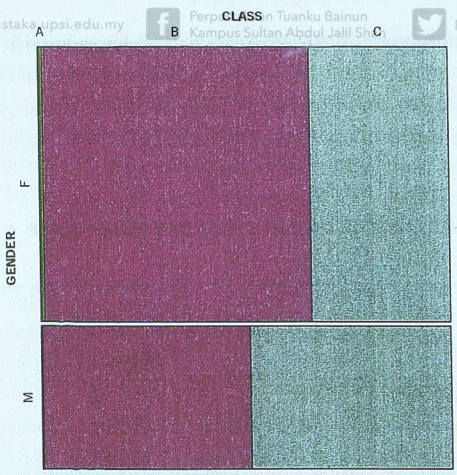


Figure 4.8: Degree classification by gender

4.2.4 Performance According to Age Group

Examination of the cell frequencies in crosstabulated degree classification by age group (Table 4.5) shows that out of 0.649 percent of all first class students, majority are students from age group U1 (29 years old or below) with 55.1 percent, followed by students in group U3 (40 - 49 years old) with 22.4 percent.

Table 4.5: Cross tabulation of age group and degree classification

			Class			Row
			First	2nd Upper	2nd Lower	Total
Age group	U1	Count	81	8237	3054	11372
		% within Age	0.712	72.432	26.855	
		% within Class	55.102	60.687	34.230	
	U2	Count	23	3018	3184	6225
		% within Age	0.369	48.482	51.149	
		% within Class	15.646	22.235	35.687	
	U3	Count	33	1821	2286	41403
		% within Age	0.797	43.986	55.217	
		% within Class	22.449	13.416	25.622	
	U4	Count	10	497	398	905
		% within Age	1.105	54.917	43.978	
		% within Class	6.803	3.662	4.461	
Column Total		Count	147	13573	8922	22642
		% within Class	0.649	59.946	39.405	100

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.5 Performance According to Race

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.6 shows an association between race and degree classification. The crosstabulated degree classification by race shown in Table 4.6 reveals that among first class students, 44.9 percent are chinese and 48.2 percent are malay. Result also shows that none of the Bumiputera Sabah and Origin native students receives first class degree.

Table 4.6: Cross tabulation of race and degree classification

			Class			
			First	2nd Upper	2nd Lower	Row Total
Race	Bumiputera Sabah	Count	**0**	391	219	610
		% within Race	0.000	64.098	35.902	
		% within Class	0.000	2.881	2.455	
	Bumiputera Sarawak	Count	1	276	190	467
		% within Race	0.214	59.101	40.685	
		% within Class	0.680	2.033	2.130	
	Chinese	Count	66	1044	173	1283
		% within Race	5.144	81.372	13.484	
		% within Class	44.898	7.692	1.939	
	Foreigner	Count	1	16	13	30
		% within Race	3.333	53.333	43.333	
		% within Class	0.680	0.118	0.146	
	Indian	Count	18	476	389	873
		% within Race	0.916	54.525	44.559	
		% within Class	5.442	3.507	4.360	
	Malay	Count	71	11322	7856	19249
		% within Race	0.369	58.819	40.813	
		% within Class	48.299	83.416	88.052	
	Origin Native	Count	**0**	1	2	3
		% within Race	0.000	33.333	66.667	
		% within Class	0.000	0.007	0.0226	
	Others	Count	**0**	47	80	127
		% within Race	0.000	37.008	62.992	
		% within Class	0.000	0.346	0.897	
Column Total		Count	147	13573	8922	22642
		% within Class	0.649	59.946	39.405	100

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.6 Performance According to Gender and Faculty

Figure 4.9 shows the performance related to gender and faculty. In general, both female and male students show an almost similar pattern of academic performance. Eighty nine percent of the first class graduates are from four faculties namely FPE, FST, FSKPM and FB and 69.3 percent of them are females. Out of 34 percent of first class graduates from PFE, 94 percent are females. The similar trends is also shown for FST, FSKPM and FB with 91.4 percent, 75 percent and 94.4 percent respectively.

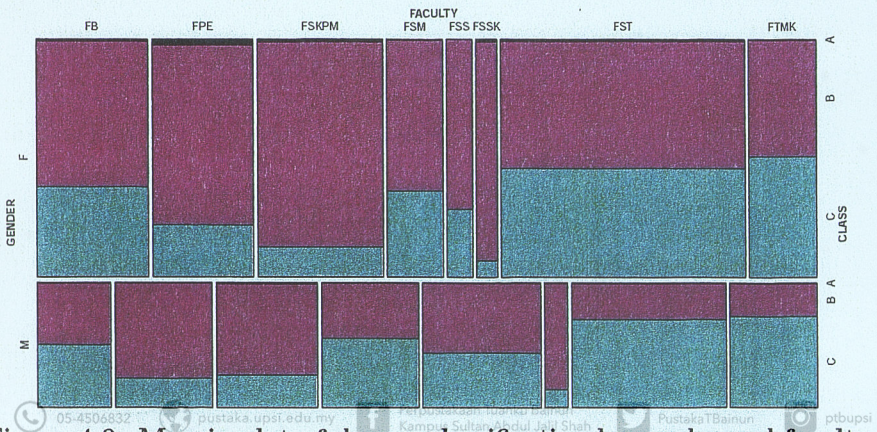


Figure 4.9: Mosaic plot of degree classification by gender and faculty

4.2.7 Performance According to Gender and Intake Category

Figure 4.10 shows the performance related to gender and qualifications at entry. The majority or 41 percent of the students are from matriculation, 26 percent are from PKPG, 10 percent are from special case, 8 percent are diploma holder, 7 percent are from second channel, 5 percent are from KDPK program and the remaining 3 percent are STPM holder. Matriculation and PKPG students are the majority group in the university with composition of 67 percent out of total students. Even though STPM students are the minority group for both genders, they outperform other groups.

Out of 88 percent of the female students who receive first class degree, 67 percent are STPM holders, 27 percent are either from matriculation or PKPG and 6 percent are diploma holders. Out of 12 percent of the male students received first class degree, they are either from diploma or STPM.

Performance for all other groups except STPM for male students are almost equal. However, the differences among female students are not consistent for various category of intake category. Matriculation students achieved at the same level as diploma students while PKPG and KDPK are slightly lower. For the female students, second channel achieved the worst.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

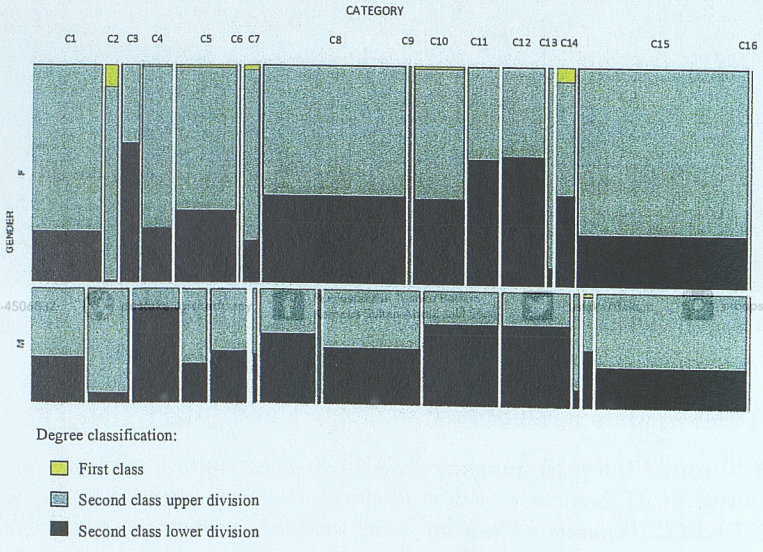


Figure 4.10: Mosaic plot of degree classification by gender and intake category

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.8 Performance According to Age Group and Race

Figure 4.11 shows the performance related to age group and race. In term of first class degree holder, malay and chinese far outperform members of other race groups. Out of 5.1 percent of chinese who receive first class degree, 89.4 percent are either in group U1 or group U3. Seventy seven percent of malay students who receive first class degree are from group U1 and group U2.

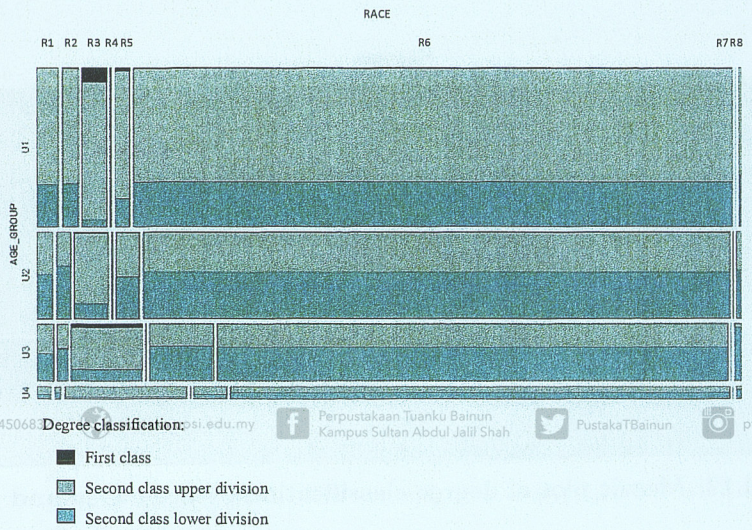


Figure 4.11: Mosaic plot of degree classification by age group and race

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.9 Performance According to Age Group and Gender

In terms of age, 47 percent of the students begin their study at this university at the age of 29 or younger. Thirty one percent are between 30 and 39, 18 percent between 40 and 49, and the remaining 4 percent are 50 or above. Figure 4.12 shows a similar pattern of academic performance for both female and male students. In general, the students in group U1 and U4 perform slightly better than the students in group U2 and U3. Out of 4 percent of students in group U4, 1.1 percent of students received first class degree. On the other hand, out of 50.2 percent students in group U1, only 0.7 percent received first class degree.

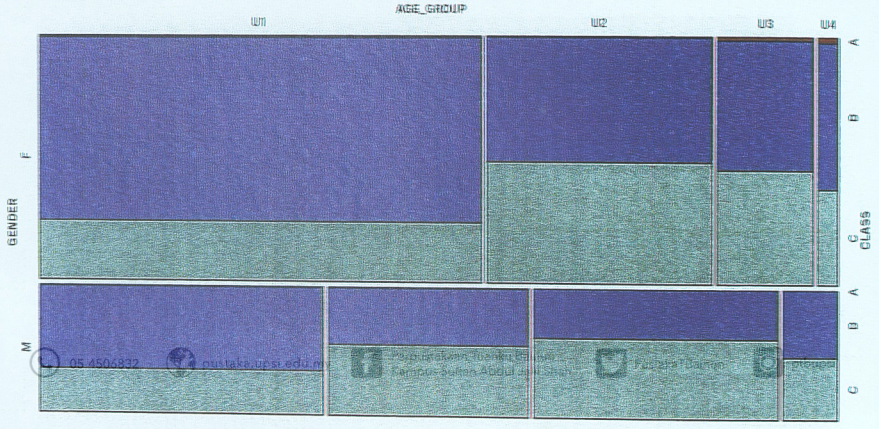


Figure 4.12: Mosaic plot of degree classification by age group and gender

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.10 Performance According to Age Group and Faculty

Figure 4.13 reveals the pattern of degree classification according to age group and faculty of study. In general, younger students perform better than older students for all faculties. However, at the FPE, student in group U3 and U4 perform better than other groups where more than 90 percent of students in these group received first class degree or second class degree upper division.

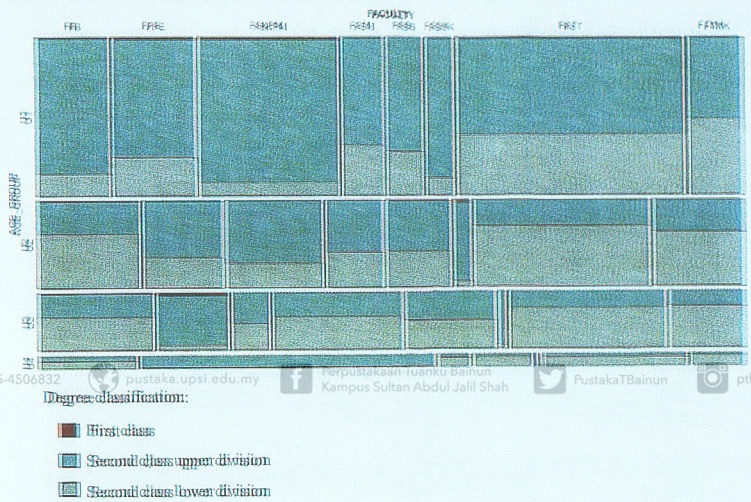


Figure 4.13: Mosaic plot of degree classification by age group and faculty

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.11 Performance According to Age Group and Intake Category

Figure 4.14 shows the pattern of degree classification according to age group and intake category. Out of 0.7 percent of students in group U1 who received first class degree, 45.7 percent are STPM holder, 25.9 percent are from special case, 17.5 percent are from matriculation and the remaining 10.9 percent are either from MPSAH or diploma. The performance in group U2 are almost similar for all intake category except for teacher certificate holder where 92.3 percent of them received second class upper division. For group U3 and group U4, students from Institute of Aminuddin Baki perform better than other categories. Seventy percent of first class degree students in group U3 are from Institute of Aminuddin Baki and out of 1.1 percent of first class degree in group U4, 90 percent are also intake from Institute of Aminuddin Baki.

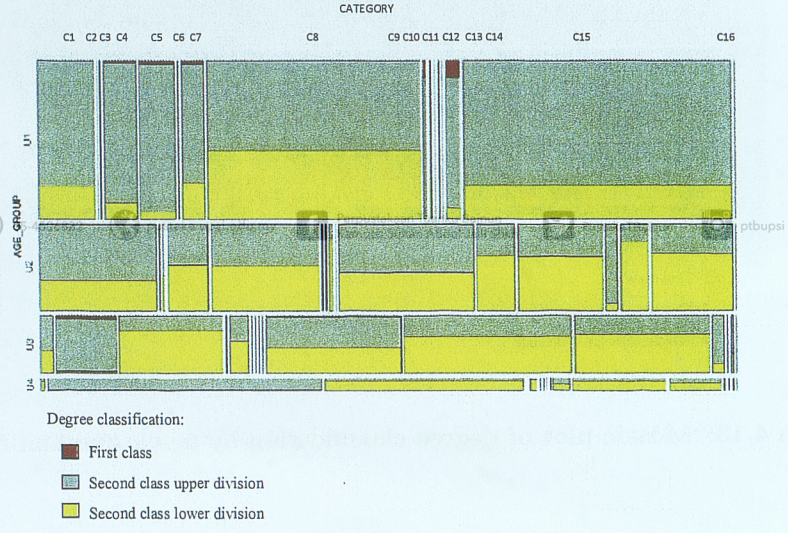


Figure 4.14: Mosaic plot of degree classification by age group and intake category

4.2.12 Performance According to Race and Faculty

Figures 4.15 and 4.16 reveal the performance for Bumiputra and Non Bumiputera students across faculty. At the FPE, the percentage of Bumiputera students who receive first class degree are equal to the Non Bumiputera students. At the FB, 33.3 percent of Bumiputera students received first class degree compared to 66.7 percent of Non Bumiputera students. The similar pattern is seen at the FST where 22.9 percent of Bumiputera students received first class degree and 77.1 percent for

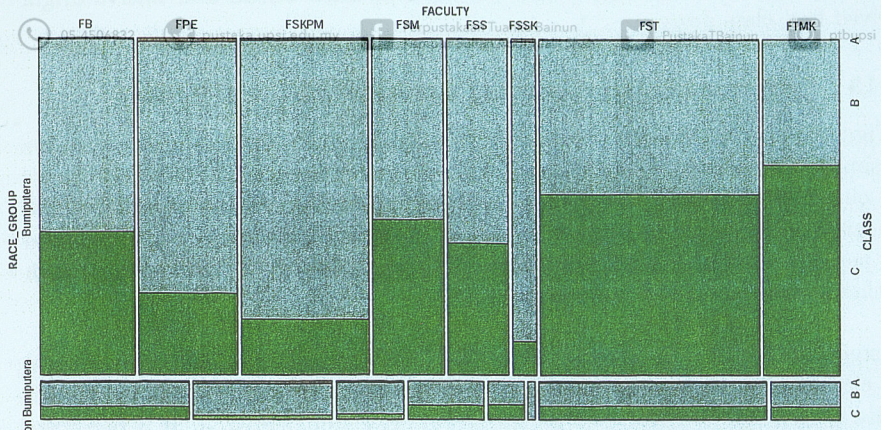


Figure 4.15: Mosaic plot of degree classification by race group and faculty

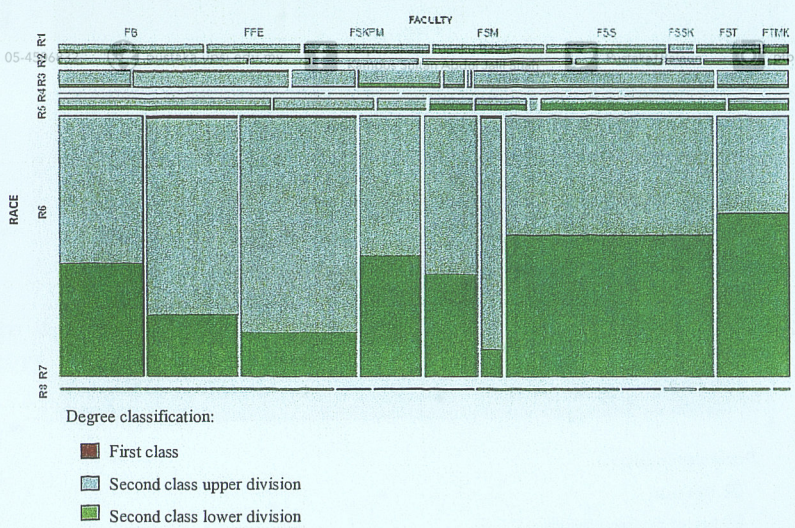


Figure 4.16: Mosaic plot of degree classification by race and faculty

Non Bumiputera students. However, a different pattern is shown at the FSKPM where 75 percent of Bumiputera received first class degree. There are none of the Non Bumiputera students received first class degree at the FSM, FSS and FSSK.

The result also revealed that, none of the Bumiputera Sabah and native origin received first class degree.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.2.13 Performance According to Race and Intake Category

In Figure 4.17, performance for Bumiputra students across intake channel shows the similar pattern as in female students shown in Figure 4.10. Figure 4.18 shows that out of 49 percent of Bumiputera students who received first class degree, 33.3 percent are from special case, 18.1 percent are from matriculation programme and 16.7 percent are from PKPG or Institute of Aminuddin Baki. Percentage for each of other channels is less than 10 percent.

As expected, out of 51 percent of the total Non-Bumiputra students who receive first class degree, 45.3 percent are STPM holders and 26.7 percent are from Institute of Aminuddin Baki. For the second class upper degree, 38.3 percent are STPM holder, followed by PKPG students with 28.9 percent and 11.1 percent from Institute of Aminuddin Baki. Second class lower degree is dominated by PKPG students with 42.4 percent out of total 7.3 percent of the Non-Bumiputra students who receive second class lower degree.



Figure 4.17: Mosaic plot of degree classification according to race group and intake category

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

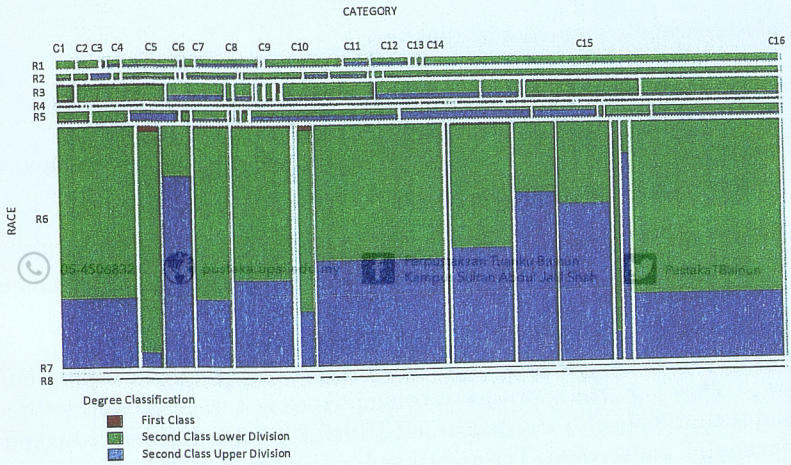


Figure 4.18: Mosaic plot of degree classification by race and intake category

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.3 Data Analysis of Academic Performance Using Meta-Analysis

Meta-analysis is a statistical analysis of a collection of individual studies proposed by Glass (1976). It is a method to integrate and summarize the findings from a collection of studies that address the similar questions. Meta-analysis can give a useful increase in precision (Cumming, 2011).

There have been many studies conducted to study the effect of gender on academic performance such as in Jabor et al. (2011), Lindberg et al. (2010), Ismail and Awang (2008) and Ding et al. (2006) but they showed conflicting results. For example, Jabor et al. (2011), Ismail and Awang (2008) and Ding et al. (2006) reported that there were statistically significant differences in academic performance between gender. On the other hand, Lindberg et al. (2010) discovered that males and females perform similarly in mathematics. No meta-analysis has yet been made of gender differences in all faculties in a university.

Meta-analysis generally proceed in three steps:

1. The researchers locate all studies related to the topic being reviewed.

2. Statistics are extracted from each report and an effect size is computed for each individual study.

3. An average of the effect sizes is computed to obtain an overall assessment of the direction and magnitude of the treatment when all studies are combined.

Forest plot is a useful graphical method of displaying treatment effects across subgroups. They are presented with two columns to give the differences between studies and the estimate of the overall result (Ried, 2006). The left-hand column lists the names of the studies. The right-hand column is a plot of the measure of effect for each of the studies and represented by a square. The width of the horizontal lines through the squares indicates the 95% confidence interval of this estimate. The area of each square is proportional to the weight of each study in the meta-analysis and the main determinant of which is sample size. The overall measure of effect is plotted as a diamond. A vertical line representing no effect is also plotted. If the confidence intervals for individual studies overlap with this line, it demonstrates that at the given confidence level, their effect sizes do not differ from no effect for the individual study. Likewise, if the diamond overlap the line of no effect the overall meta-analysed result cannot be concluded to differ from no effect at the given level of confidence.

Funnel plot is a visual tool that is used to investigate publication and other bias in meta-analysis. A funnel plot is a scatter plot of the treatment effects estimated from individual studies against a measure of study size. The existence of bias is shown by an asymmetrical appearance of the funnel plot (Sterne and Harbord, 2004).

In this chapter, we analyze the students' academic performance using meta-analysis. We present the meta-analysis of gender differences in students' performance in university that is measured by their final degree classification. The results of the study provide evidence of gender differences in students' performance that is measured by their final degree classification.

In every meta-analysis, all result from each individual study need to be combined and an effect size index has to be selected to translate the results of every study into a common metric. An effect size is a measure that describes the magnitude of the difference between two groups. Effect sizes are particularly valuable in best practices research because they represent a standard measure by which all outcomes can be assessed (O'keefe and Haie, 2001).

In this meta-analysis study, we consider all eight faculties in Universiti Pendidikan Sultan Idris as an independent study. The study sample size ranging from 653 to 6600 and all studies report the odds ratio for gender difference in performance. Therefore, the odds ratio is used to represent the effect size of individual study. Odds ratio vary from 0 to infinity, with 1 indicates no effect between two binary variables. The function `rmeta()` in R software is used to conduct the meta-analysis. Table 4.7 shows the descriptive Statistics of study samples used in the meta-analysis.

Table 4.7: Descriptive statistics of male and female students in eight faculties

	Sample	First Class		2nd Class Upper		2nd Class Lower	
Faculty	Size	Female	Male	Female	Male	Female	Male
FB	2911	17	1	1323	370	822	378
FPE	2920	47	3	1469	749	425	227
FSKPM	3470	21	7	2113	758	306	265
FSM	2089	2	1	698	439	398	551
FSS	1724	0	3	358	684	143	536
FSSK	653	4	1	377	207	29	35
FST	6600	32	3	2646	478	2272	1169
FTMK	2275	2	3	669	244	692	674

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.4 Meta-Analysis for First Class Degree Classification

Table 4.8 illustrates that an independent odds ratio test reveals female students are more likely to obtain first class degree than their male counterparts except for faculty FTMK. Note that FSS is not estimated since there is no events in female group. Such studies will be excluded from the meta-analysis.

Table 4.8: Meta-analysis for first class degree between female and male students

Faculty	Female		Male		OR	CI
	1st Class	Not 1st Class	1st Class	Not 1st Class		
FB	17	2145	1	748	5.93	(0.79, 44.62)
FPE	47	1894	3	976	8.07	(2.51, 26.00)
FSKPM	21	2419	7	1023	1.27	(0.54, 2.99)
FSM	2	1096	1	990	1.81	(0.16, 19.95)
FSS	0	501	3	1220	0.00	(0.00, Not Available)
FSSK	4	406	1	242	2.38	(0.26, 21.46)
FST	32	4918	3	1647	3.57	(1.09, 11.68)
FTMK	2	1352	3	918	0.45	(0.08, 2.71)
Overall Effect					2.76	(1.72,4.42)

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Figure 4.19 shows the forest plot of the effect size of eight studies. The vertical line in the middle (OR = 1) is known as the line of no effect. The black squares represent the odds ratios of the individual studies. The horizontal lines through the squares represent their 95% confidence intervals (CI). The width of the CI is related to the power of the study, the wider the CI, the less precise the study results. if the CI crosses the vertical line of no effect, we cannot reject the null hypothesis of no difference between the two groups. The size of the black squares reflects the weight of each study in the meta-analysis. The weight of each study is a function of study size. The diamond in the last row of the graph illustrates the pooled or overall result of the meta-analysis and its width is the CI for the pooled data.

In the forest plot shown in Figure 4.19, the result of eight studies for first class degree have been pooled. Study from faculties FB, FSM, FSSK and FTMK are considered as small studies, with a wide confidence interval and crossing the line of no effect (OR = 1). On the other hand, Faculty FSKPM is a larger study with narrow confidence interval but cross the line of no effect. Therefore, we are unable to conclude if the female students in those faculties are more likely to obtain first class degree than male students. Study from faculties FPE and FST are considered as larger study and do not cross the line of no effect. This proves a weak evidence that the female students in those faculties are more likely to obtain first class degree. Across all 8 studies, the weighted mean odds ratio is 2.76. The lower and upper bounds of 95% confidence for the true population odds of female

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

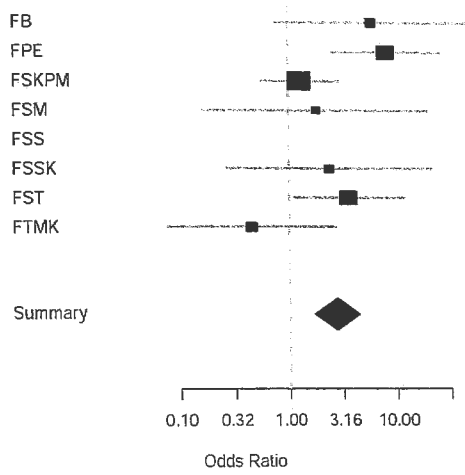


Figure 4.19: Forest plot of observed effect sizes and the 95% confidence intervals for the first class degree classification studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

students obtained first class degree are 1.72 and 4.42.

Figure 4.20 shows the funnel plot of meta-analysis for first class degree classification. The symmetrical plot shown in Figure 4.20 indicates the absence of bias in the studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

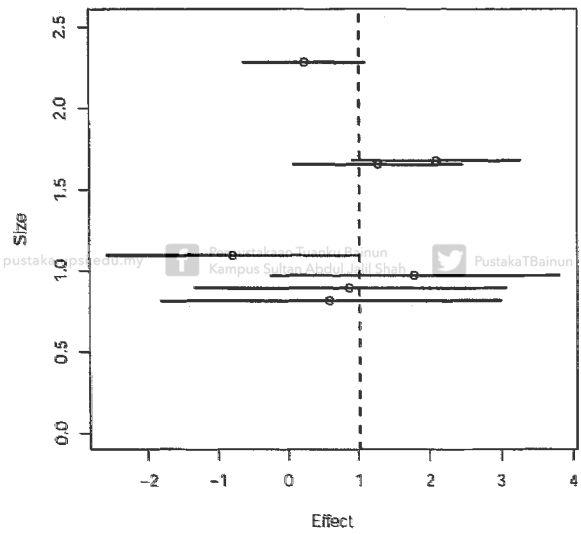


Figure 4.20: Funnel plots for the first class degree classification studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.5 Meta-Analysis for Second Class Upper Degree Classification

Table 4.9 illustrates that an independent odd ratio test reveals female students are more likely to obtain second class upper degree than their male counterparts except for faculty FPE.

Table 4.9: Meta-analysis for second class upper degree between female and male students

Faculty	Female		Male		OR	CI
	2nd Class Upper	Not 2nd Class Upper	2nd Class Upper	Not 2nd Class Upper		
FB	1323	839	370	379	1.62	(1.37, 1.91)
FPE	1469	472	749	230	0.96	(0.80, 1.14)
FSKPM	2113	327	758	272	2.32	(1.93, 2.78)
FSM	698	400	439	552	2.19	(1.84, 2.62)
FSS	358	143	684	539	1.97	(1.58, 2.47)
FSSK	377	33	207	36	1.99	(1.20, 3.28)
FST	2646	2304	478	1172	2.82	(2.50, 3.18)
FTMK	660	694	244	677	2.64	(2.20, 3.16)
Overall Effect					2.08	(1.95,2.21)

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

In the forest plot shown in Figure 4.21, the result of eight studies for second class upper degree have been pooled. Study from faculties FSSK is considered as small studies while FSK is considered as largest study. All studies except FPE lie on the right line of no effect without crossing it. This proves that female students in faculties FB, FSKPM, FSM, FSS, FSSK, FST and FTMK are more likely to obtain second class upper degree than male students. Faculty FPE has narrow confidence interval but crosses the line of no effect. Therefore, we are unable to conclude if the female students in FPE are more likely to obtain second class upper degree than male students. Across all 8 studies, the weighted mean odds ratio is 2.08. The lower and upper bounds of 95% confidence for the true population odds of female students obtained second class upper degree are 1.95 and 2.21.

Figure 4.22 shows the funnel plot of meta-analysis for second class upper degree classification. The symmetrical plot shown in Figure 4.22 indicates the absence of bias in the studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

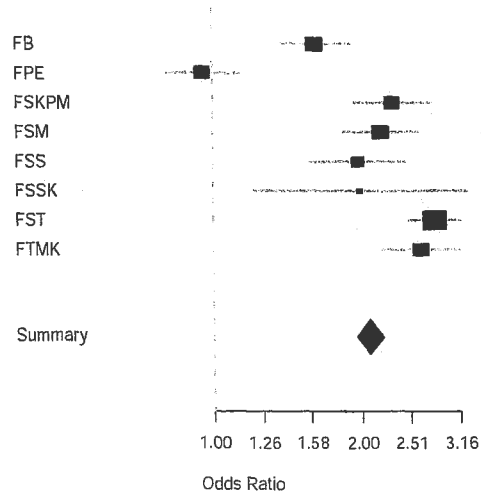
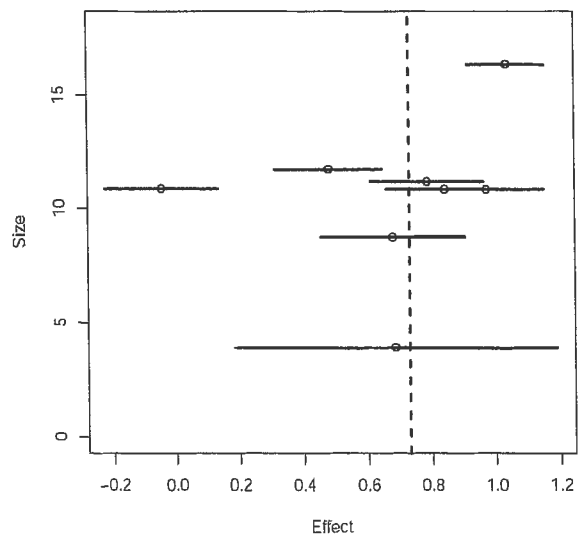


Figure 4.21: Forest plot of observed effect sizes and the 95% confidence intervals for the second class upper degree classification studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi



05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Figure 4.22: Funnel plots for the second class upper degree classification studies.

4.6 Meta-Analysis for Second Class Lower Degree Classification

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.10 illustrates that an independent odd ratio test reveals female students are less likely to obtain second class lower degree than their male counterparts.

Table 4.10: Meta-analysis for second class lower degree between female and male students

Faculty	Female		Male		OR	CI
	2nd Class Lower	Not 2nd Class Lower	2nd Class Lower	Not 2nd Class Lower		
FB	822	1340	378	371	0.60	(0.51, 0.71)
FPE	425	1516	227	752	0.93	(0.77, 1.12)
FSKPM	306	2134	265	765	0.41	(0.34, 0.50)
FSM	398	700	551	440	0.45	(0.38, 0.54)
FSS	143	358	536	687	0.51	(0.41, 0.64)
FSSK	29	381	35	208	0.45	(0.27, 0.76)
FST	2272	2678	1169	481	0.35	(0.31, 0.39)
FTMK	692	662	674	247	0.38	(0.32, 0.46)
Overall Effect					0.47	(0.44,0.50)

In the forest plot shown in Figure 4.23, the result of eight studies for second class lower degree have been pooled. Study from faculties FSSK is considered as small studies while FSK is considered as the largest study. All studies except FPE lie on the left without crossing the line of no effect. This prove that female students in faculties FB, FSKPM, FSM, FSS, FSSK, FST and FTMK are less likely to obtain second class lower degree than male students. Faculty FPE has a narrow confidence interval but cross the line of no effect. Therefore, we are unable to conclude if the female students in FPE are less likely to obtain second class lower degree than male students. Across all 8 studies, the weighted mean odds ratio is 0.47. The lower and upper bounds of 95% confidence for the true population odds of female students obtained second class lower degree are 0.44 and 0.5.
Figure 4.24 shows the funnel plot of meta-analysis for second class lower degree classification. The symmetrical plot shown in Figure 4.24 indicates the absence of bias in the studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

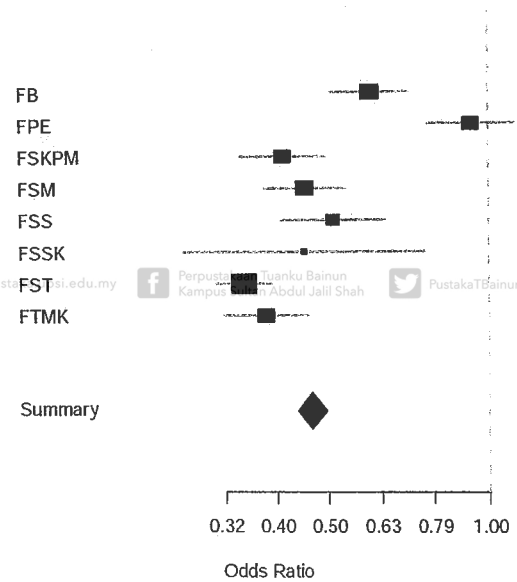


Figure 4.23: Forest plot of observed effect sizes and the 95% confidence intervals for the second class lower degree classification studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

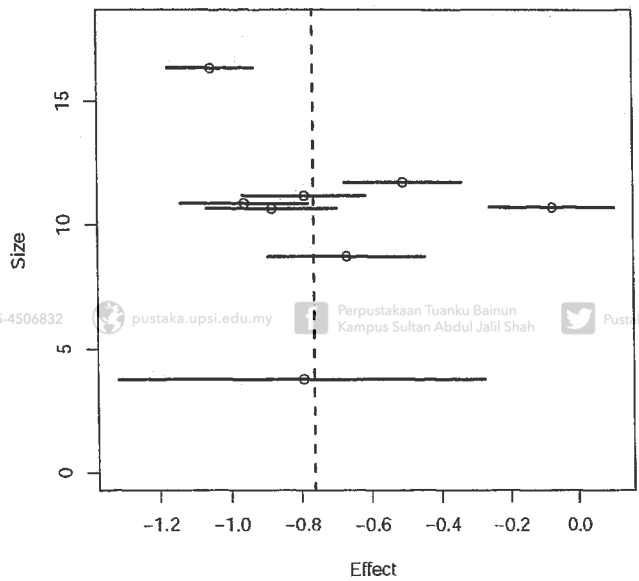


Figure 4.24: Funnel plots for the second class lower degree classification studies.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

The classification tree generated from students' background data is shown in Figure 4.25. The classification accuracy of the model is 73.88 percent. This would suggest that the students' background information gathered during the enrolment process contain sufficient information for early identification of successful and less successful students. The binary classification tree produces six levels and has seventeen nodes including nine leaf nodes. There are four predictors, namely faculty, intake category, age group and race influenced the final degree classification of students. The description of the code used in the classification tree can be found in the Appendix A.

The root node had 22,642 observations: class A with 147 observations, class B with 13,573 observations and class C with 8,922 observations. The first split is based on the FACULTY. The left branch, FACULTY F2, F3 and F6 contains 7,043 observations and the right branch, FACULTY F1, F4, F5, F7 and F8 contains 15,599 observations. The left branch is split into two terminal nodes based on intake category: CATEGORY C1, C2, C4, C5, C7, C10, C12, C13, C15 and C16 is classified as class B having 5,573 observations and CATEGORY C3, C6, C8 and C14 is classified in CLASS C having 219 observations. The right branch has more splits based on AGE GROUP, CATEGORY, RACE and FACULTY.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

The standard tree model is unable to classify the degree classification for class A because its percentage is too small. To overcome this problem, we display the probability for each degree classification at all leaf nodes. For example, the probability for students classified in the left most leaf node to obtain first class, second class upper and second class lower degree is 0.012, 0.829 and 0.159 respectively.
The result of CART identifying the importance of predictor variables that affect the student degree classification. Obviously, faculty is the main variable used in this classification, followed by intake category, age group and race. Gender is not represented in the tree model. This indicates that gender is not an important factor to classify sudents' academic performance. Age group is an important factor to classify students in FB, FSM, FSS, FST and FTMK while race is important to classify the performance of older students. The students' academic performance in faculty F2, F3 and F6 are only influenced by their intake category. It also shows that there are no class A students in faculty F1, F4 and F8 for the group of students entering university at the age below 29 years old. It is also interesting to note that race only influence the performance of older students. Nine classification rules produced by this tree model according to branches are shown in Table 4.11 to Table 4.13.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

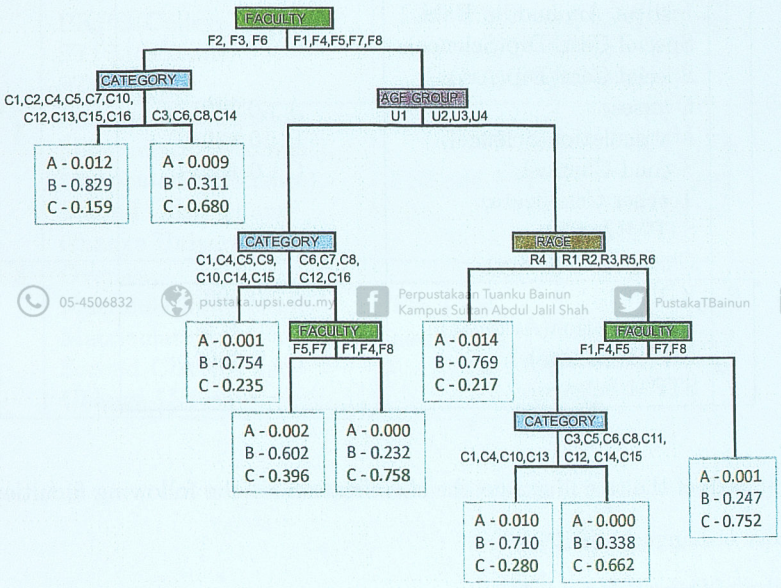


Figure 4.25: Classification tree model of students degree classification

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.11 shows the classification rules for the left branch which include the following faculties:

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

- Faculty of Management and Economics (FPE)
- Faculty of Cognitive Science and Human Development (FSKPM)
- Faculty of Human Science (FSSK)

Table 4.11: Classification Rules for the Left Branch

	Intake Category	Degree Classification
1	Diploma, Institut Aminuddin Baki, Special Case/Dip(Science), Special Case/Dip(Art), Oversea, Matriculation(Science), Second Channel, Teacher Certificate, STPM(Art), Change University,	A : 0.012 B : 0.829 C : 0.159
2	KDPK , Matriculation(Account), PKPG(College) STPM(Science)	A : 0.009 B : 0.311 C : 0.680

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

The right branch of the tree illustrate the classification for the following faculties:

- Faculty of Language (FB)
- Faculty of Sport Science (FSS)
- Faculty of Music (FSM)
- Faculty of Science and Technology (FST)
- Faculty of Information Technology and Communication (FTMK)

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

This right branch divides the students according to their age group. The classification for the students entering the university at the age 29 or below is given in Table 4.12. The classification for the students entering the university at the age 30 and above is given in Table 4.13.

Table 4.12: Classification Rules for the students entering the university at the age 29 or below

	Intake Category	Faculty	Degree Classification
1	Diploma Special Case/Dip(Science) Special Case/Dip(Art) MPSAH PKPG(College) STPM(Science) STPM(Art)	-	A : 0.001 B : 0.754 C : 0.235
2	Oversea Qualification, Matriculation(Account), Matriculation(Science), Second Channel and Change University	FSS and FST	A : 0.002 B : 0.602 C : 0.396
3	Oversea, Matriculation(Account), Matriculation(Science), Second Channel and Change University	FB, FSM and FTMK	A : 0.000 B : 0.232 C : 0.758

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.13: Classification rules for the students entering the university at the age 30 or above

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

	Race	Faculty	Intake Category	Degree Classification
1	Chinese	-	-	A : 0.014 B : 0.769 C : 0.217
2	Bumiputera Sabah, Bumiputera Sarawak, Indian, Chinese, Malay and Native Origin	FB, FSM and FSS	Diploma, Special Case/ Dip(Science), PKPG(College) and Teacher Certificate	A : 0.010 B : 0.710 C : 0.280
3	Bumiputera Sabah, Bumiputera Sarawak, Indian, Chinese, Malay and Native Origin	FB, FSM and FSS	KDPK, Special Case/Dip(Art), Oversea Matriculation(Science) PKPG(UPSI) Second Channel STPM(Science) STPM(Art)	A : 0.000 B : 0.338 C : 0.662
4	Bumiputera Sabah, Bumiputera Sarawak, Indian, Chinese, Malay and Native Origin	FST and FTMK	-	A : 0.001 B : 0.247 C : 0.752

The naive Bayes structure is shown in Figure 4.26. The Bayesian network structure in Figure 4.27 shows relationships between the degree classification with students' background data. It reveals that the degree classification has strong dependency on intake category and faculty. Each node is described by a conditional probability distribution on its direct predecessors.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

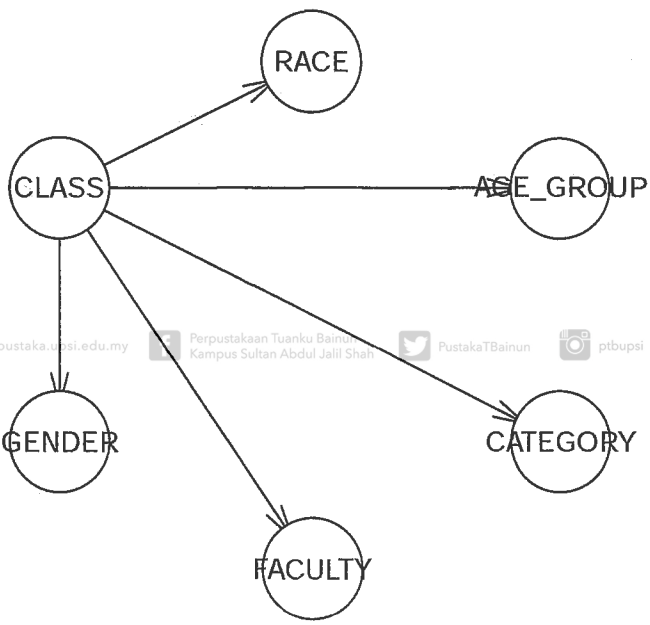


Figure 4.26: Naive Bayes classification model of students degree classification

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

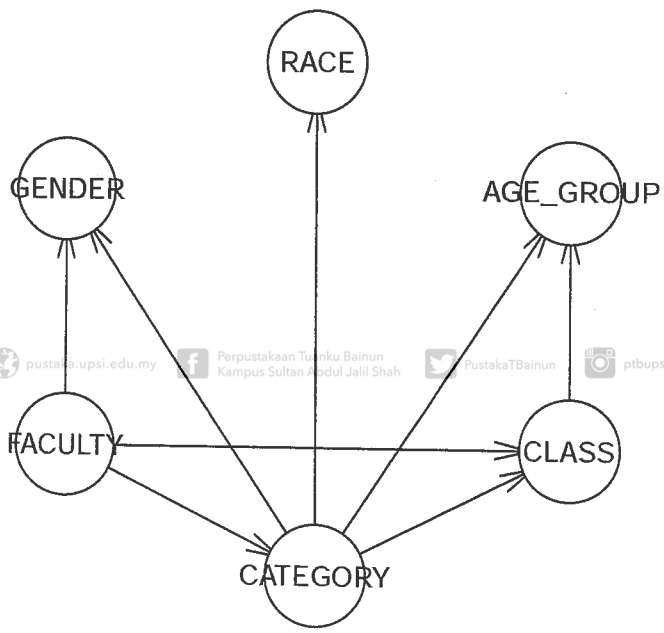


Figure 4.27: Bayesian network classification model of students degree classification

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

As we can see in the Table 4.14, classification tree perform the best with accuracy 73.88 percent, followed by Bayesian network with 71.23 percent and naive Bayes perform the worse with accuracy 69.31 percent.

Table 4.14: Classification accuracy for classification tree, bayesian network and naive bayes

Dataset	Classification tree	Bayesian network	Naive Bayes
Students' background data	73.88%	71.32%	69.31%

Tables 4.15 and 4.16 show the confusion matrix for three classes prediction using classification tree. It shows that all 147 first class students are incorrectly classified using classification tree. Out of 13573 second class upper degree, 11479 students are correctly classified. So, the true positive rate for second class upper degree is 84.6 percent and the false positive rate is 42.0 percent. Out of 8922 second class lower degree, 5250 students are correctly classified. So, the true positive rate for second class lower degree is 58.8 percent and the false positive rate is 15.3 percent.

Table 4.15: Confusion matrix for degree classification using classification tree

		Predicted Class		
		First Class	2nd Class Upper	2nd Class Lower
Actual Class	First Class	0	141	6
	2nd Class Upper	0	11479	2094
	2nd Class Lower	0	3672	5250

Table 4.16: Class wise accuracy for three classes prediction using classification tree

Class	True Positive Rate	False Positive Rate
First Class	0	0
2nd Class Upper	0.846	0.420
2nd Class Lower	0.588	0.153

Tables 4.17 and 4.18 show the confusion matrix for three classes prediction using Bayesian network. As can be observed, Bayesian network is also unable to correctly classify all 147 first class students. Out of 13573 second class upper degree, 11279 students are correctly classified. So, the true positive rate for second class upper degree is 83.1 percent and the false positive rate is 46.2 percent. Out of 8922 second class lower degree, 4865 students are correctly classified. So, the true

positive rate for second class lower degree is 45.5 percent and the false positive rate is 16.8 percent.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.17: Confusion matrix for degree classification using Bayesian network

		Predicted Class		
		First Class	2nd Class Upper	2nd Class Lower
Actual Class	First Class	0	135	12
	2nd Class Upper	0	11279	2294
	2nd Class Lower	0	4057	4865

Table 4.18: Class wise accuracy for three classes prediction using Bayesian network

Class	True Positive Rate	False Positive Rate
First Class	0	0
2nd Class Upper	0.831	0.462
2nd Class Lower	0.455	0.168

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Tables 4.19 and 4.20 show the confusion matrix for three classes prediction using naive Bayes. As can be observed, Out of 147 first class degree, 26 students are correctly classified. So, the true positive rate for first class degree is 17.7 percent and the false positive rate is 0.5 percent. Out of 13573 second class upper degree, 10703 students are correctly classified. So, the true positive rate for second class upper degree is 78.9 percent and the false positive rate is 34.1 percent. Out of 8922 second class lower degree, 4953 students are correctly classified. So, the true positive rate for second class lower degree is 55.5 percent and the false positive rate is 15.7 percent.

Table 4.19: Confusion matrix for degree classification using naive Bayes

		Predicted Class		
		First Class	2nd Class Upper	2nd Class Lower
Actual Class	First Class	26	104	17
	2nd Class Upper	115	10703	2755
	2nd Class Lower	1	3968	4953

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

This results show that for large categorical dataset, classification tree perform the best in classification accuracy, followed by Bayesian network and naive Bayes. However, both classification tree and bayesian network are unable to classify first

Table 4.20: Class wise accuracy for three classes prediction using naive Bayes

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Class	True Positive Rate	False Positive Rate
First Class	0.177	0.005
2nd Class Upper	0.788	0.341
2nd Class Lower	0.555	0.157

class degree. Although naive Bayes produces the least accuracy, it is able to classify first class degree. Thus, we conclude that classification tree model is acceptable and good enough to serve as an early identification of student performance.

4.8 Simulation of Population Data

The simulation data is used to model the dataset using classification tree, Bayesian network and naive Bayes for three purposes. Firstly it it use to compare the prediction of students' academic performance using. Secondly, to investigate the influence of different missing data mechanism and finally to impute the missing values. These three processes are carried out in order to select the best model and imputation method to accurately predict students' academic performance based on their background data with the present of missing values. The simulation process consider randomness through properly identified probability distributions taken directly from real data.

According to Munnich and Schurle (2003), the following conditions need to be considered when simulating population data:

1. The concrete sizes of regions and strata should be considered precisely.

2. The heterogeneity of the data distributions should reflect the real settings including cross correlations on region and stratum level.

3. Marginal distributions of all variables should be considered as they were known.

4. Disclosure control rules must not be violated.

Generating multivariate data involves the specification of correlation between covariates. We apply the method proposed by Ferrari and Barbiero (2012). The data are generated in two seps:

1. Sets up the original continuous variables in order to achieve the final discrete or ordinal variables meeting the experimental conditions.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

2. Generates samples from the adjusted original variables and discretizes them, thus simulating samples from the target variables.

4.8.1 Algorithm for Simulation of Population Data

We generated 22642 synthetic categorical data datasets with six attribute variables and three class labels with pre-specified correlation matrix and marginal distributions. Data are generated separately for different degree classifications.

- STEP 1
  Calculate the marginal of population data.
  Set the marginal based on the population data.
- STEP 2 Check the lower bound and the upper bound of the correlation coefficients.
- STEP 3 Sets the correlation coefficients. The values must be between the lower bound and the upper bound of the correlation coefficients.

The correlation matrix and marginal distributions are calculated using the undergraduate students data obtained from the Universiti Pendidikan Sultan Idris (UPSI) who's graduated from 1997 to 2008. The simulation process consider randomness through properly identified probability distributions taken directly from study data.
The correlation matrix first class degree dataset is as follows:

$$\begin{pmatrix} 1.0 & 0.63 & 0.69 & 0.42 & 0.55 & 0.32 \\ 0.63 & 1.0 & 0.94 & 0.81 & 0.93 & 0.24 \\ 0.69 & 0.94 & 1.0 & 0.79 & 0.90 & 0.29 \\ 0.42 & 0.81 & 0.79 & 1.0 & 0.91 & 0.15 \\ 0.55 & 0.93 & 0.90 & 0.91 & 1.0 & 0.18 \\ 0.32 & 0.24 & 0.29 & 0.15 & 0.18 & 1.0 \end{pmatrix}$$

The correlation matrix second class upper degree dataset is as follows:

$$\begin{pmatrix} 1.0 & 0.83 & 0.83 & 0.26 & 0.76 & 0.21 \\ 0.83 & 1.0 & 0.86 & 0.54 & 0.97 & 0.27 \\ 0.83 & 0.86 & 1.0 & 0.29 & 0.77 & 0.36 \\ 0.26 & 0.54 & 0.29 & 1.0 & 0.62 & 0.28 \\ 0.76 & 0.97 & 0.77 & 0.62 & 1.0 & 0.23 \\ 0.21 & 0.27 & 0.36 & 0.28 & 0.23 & 1.0 \end{pmatrix}$$

The correlation matrix second class lower degree dataset is as follows:

$$\begin{pmatrix} 1.0 & 0.74 & 0.78 & 0.26 & 0.80 & 0.11 \\ 0.74 & 1.0 & 0.85 & 0.52 & 0.95 & 0.23 \\ 0.78 & 0.85 & 1.0 & 0.36 & 0.91 & 0.15 \\ 0.26 & 0.52 & 0.36 & 1.0 & 0.53 & 0.61 \\ 0.80 & 0.95 & 0.91 & 0.53 & 1.0 & 0.23 \\ 0.11 & 0.23 & 0.15 & 0.61 & 0.23 & 1.0 \end{pmatrix}$$

Figure 4.28 and Figure 4.29 show the mosaic plots generated using real dataset and simulated dataset of student's background. Both real dataset and simulated

Table 4.21: Summary of simulation data

Variable		Percentage
Gender	G1	65.34
	G2	34.66
Race	R1	84.76
	R2	5.72
	R3	4.14
	R4	2.76
	R5	1.98
	R6	0.09
	R7	0.12
	R8	0.45
Age Group	U1	50.60
	U2	27.30
	U3	18.15
	U4	3.95
Faculty	F1	13.45
	F2	12.85
	F3	15.40
	F4	9.13
	F5	7.72
	F6	2.87
	F7	28.51
	F8	10.07
Intake Category	C1	10.12
	C2	3.42
	C3	4.38
	C4	4.37
	C5	7.73
	C6	0.08
	C7	1.78
	C8	17.00
	C9	0.17
	C10	9.50
	C11	6.90
	C12	7.43
	C13	0.69
	C14	2.38
	C15	24.05
Degree Classification	A	0.94
	B	60.33
	C	38.73

dataset show the similar pattern of student degree classification according to faculty.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

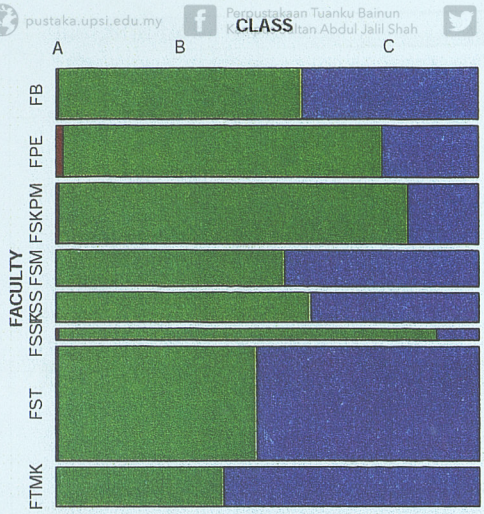


Figure 4.28: Mosaic plot of students Degree Classification using real dataset

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

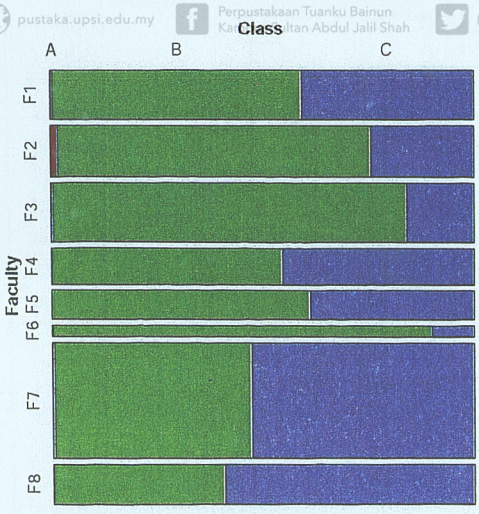


Figure 4.29: Mosaic plot of students Degree Classification using simulation dataset

4.9 Missing Data Injection

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Let the full data $\mathbf{Y} = (y_{ij})$ denotes a dataset composed of two components $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$, where $\mathbf{Y}_{obs}$ indicates completely observed sample and $\mathbf{Y}_{mis}$ indicates

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

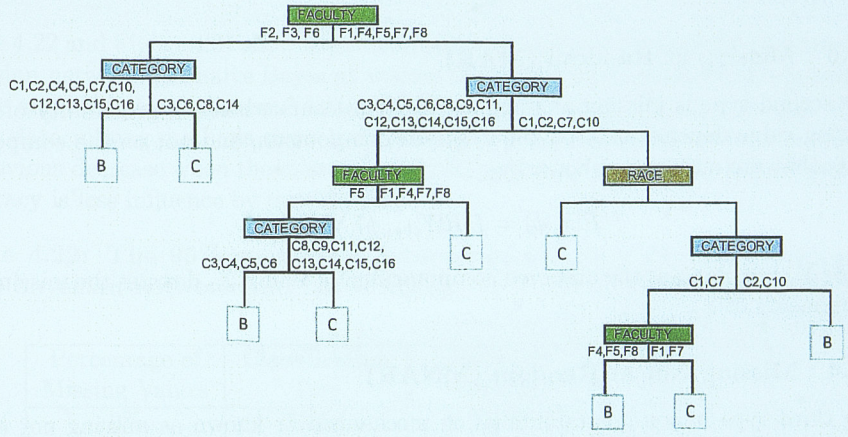


Figure 4.30: Classification tree model of students degree classification using simulation data

potentially missing sample. In other words, $\mathbf{Y}$=($\mathbf{Y}_{obs}$,$\mathbf{Y}_{mis}$). Define also the response indicators R_{ij}, such that $R_{ij} = 1$ if Y_{ij} is missing and $R_{ij} = 0$ if Y_{ij} is present. The missing data mechanism is characterized by the conditional distribution of $\mathbf{R}$ given $\mathbf{Y}$, $f(\mathbf{R}|\mathbf{Y},\phi)$, where ϕ denotes unknown parameters.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.9.1 Missing Data Mechanism

Missing data mechanisms are important since the properties of missing data methods depend closely on the nature of these mechanisms. The vital role of the mechanism in data analysis with missing values was ignored until the concept was formalized in Rubin (1976). The impact of missing data depends on the pattern of the missing data and the mechanisms that leading to missing data. According to Little and Rubin (2002), there are various type of missing data mechanism.

- Missing Completely at Random (MCAR),
- Missing at Random (MAR),
- Missing Not at Random (MNAR).

4.9.2 Missing Completely at Random (MCAR)

The first type of missing data mechanism defined as missing completely at random (MCAR) arises when the probability of a missing value for a variable is not related to the variable's value itself or to any other variable in the data set. In other word, R is independent of Y_{obs} and Y_{mis} as shown by

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

$$f(R|Y,\phi) = f(R|\phi), \forall Y, \phi,$$

where $f(R|Y,\phi)$ is conditional distribution of R given Y and ϕ denotes unknown parameters.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.9.3 Missing at Random (MAR)

The second type is missing at random (MAR), occurs when the probability of a missing value depends only on the observed components, and not on the components that are missing as shown by

$$f(R|Y,\phi) = f(R|Y_{obs},\phi), \forall Y_{mis},\phi,$$

where Y, Y_{obs} denotes the observed components of Y and Y_{mis} denotes the missing components of Y.

4.9.4 Missing Not at Random (MNAR)

The third type is non-ignorable missing mechanism or known as missing not at random (MNAR) (Diggle and Kenward, 1994), if the response probabilities depend on missing variables as shown by

$$f(R|Y,\phi) = f(R|Y_{obs},Y_{mis},\phi), \forall\phi,$$

The simulation data are used to study how classification tree is influenced by different level of missingness. For each level of missing data, we simulate missing values in one predictor variable as follows:

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

- Missing values in variable FACULTY is independent of other variables (MCAR).
- Missing values in variable FACULTY is related to variable CATEGORY (MAR).
- Missing values in variable FACULTY is related to variable FACULTY (MNAR).

Variable FACULTY is chosen to maximize the possible effect of missing values because FACULTY is the first variable used to split the dataset as shown in Figure 4.25. For MNAR, the mechanism gave 80 percent probability of missing FACULTY for FST than other faculties.

We investigate further for MCAR case where for each level of missing data, we also simulate missing values in multiple predictor variables. For missing values in multiple predictor variable, we simulate missing values as follow:

- FACULTY and CATEGORY.
- FACULTY, CATEGORY and AGE GROUP.
- FACULTY, CATEGORY, AGE GROUP and RACE.

To evaluate the influence of missing values, we randomly eliminate five levels of missing values; 5 percent, 10 percent, 15 percent, 25 percent and 50 percent. Such process is repeated 1000 times and 1000 bootstrap samples are generated for each level of missing values.

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

Table 4.22 and Figure 4.31 show the performance comparison for classification tree, bayesian network and naive Bayes at various level of missing values when applied to the actual student's dataset. As the missing rate increased, the accuracies of all three data mining methods show a decreasing pattern. Bayesian network shows an obvious decrease when the missing rate is 10 percent or more while naive Bayes accuracy is less influence by missing data.

Table 4.22: The 95% confidence interval of accuracy for classification tree, bayesian network and naive Bayes

Percentage of Missing Values	Classification Tree	Bayesian network	Naive Bayes
0%	73.88	71.32	69.31
5%	(73.43, 73.55)	(70.02, 71.32)	(68.65, 69.92)
10%	(72.80, 73.07)	(66.62, 68.06)	(68.21, 69.28)
15%	(71.98, 72.28)	(65.95, 66.97)	(68.30, 69.28)
25%	(70.42, 70.89)	(66.33, 67.26)	(68.18, 69.12)
50%	(68.99, 68.80)	(65.99, 67.46)	(68.45, 69.68)

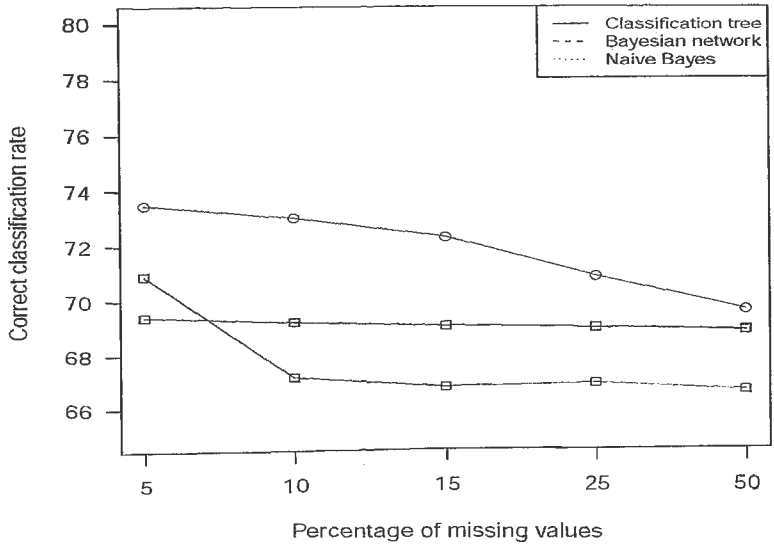


Figure 4.31: Percentage of correct classification rate for classification tree, bayesian network and naive Bayes with different level of missing values

4.11 Sensitivity of Missing Value in Classification Tree using Simulated Dataset

In this section, we summarize the standard classification tree and classification tree with missing values in Tables 4.23, 4.24, 4.25 and 4.26. Results show that splitting variable changed when the percentage of the missing variable is 10 percent or more. The classification tree is able to maintain its structure if the percentage of missing variable is not exceeded 10 percent.

	0%	5%	10%	15%	25%	50%
Predictor variables Change?	-	No	No	No	No	No
Tree Structure Change?	-	No	No	Yes	Yes	Yes
Nodes Change?	-	No	Yes	Yes	Yes	Yes

Table 4.23: Summary of the tree models at different level of missing values when missing values occur in variable FACULTY

	0%	5%	10%	15%	25%	50%
Predictor variables Change?	-	No	No	No	No	No
Tree Structure Change?	-	No	No	No	No	Yes
Nodes Change?	-	No	Yes	Yes	Yes	Yes

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.24: Summary of the tree models at different level of missing values when missing values occur in variables FACULTY and CATEGORY

	0%	5%	10%	15%	25%	50%
Predictor variables Change?	-	No	No	No	No	No
Tree Structure Change?	-	No	No	No	No	Yes
Nodes Change?	-	No	Yes	Yes	Yes	Yes

Table 4.25: Summary of the tree models at different level of missing values when missing values occur in variables FACULTY, CATEGORY and AGE GROUP

	0%	5%	10%	15%	25%	50%
Predictor variables Change?	-	No	No	No	No	No
Tree Structure Change?	-	No	No	No	No	No
Nodes Change?	-	No	Yes	Yes	Yes	Yes

Table 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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.27 shows the 95% confidence intervals for the classification tree performance of surrogate splits at various level of missing values. As the missing rate increased, the accuracies of all the classification trees for the samples are slightly decreasing. Result also shows that if the missing values for each variables is 15% or more, the tree structure will changed. Classification trees that maintain the structure shows a small reduction in classification accuracy. Likewise, classification trees that produce different tree structures also show a small reduction in accuracy. This shows that surrogate splits are able to handle MCAR for this sample.

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

Percentage of Missing Values	Number of Predictor Variable Missing				
	0	1	2	3	4
0%	85.88	-	-	-	-
5%	-	(84.03, 85.34)	(84.89, 86.27)	(85.81, 87.31)	(85.58 86.43)
10%	-	(82.56, 84.06)	(86.23, 87.32)	(86.47, 87.94)	(84.23 85.66)
15%	-	(83.17, 84.41)	(86.69, 88.00)	(86.44, 88.27)	(84.01, 85.65)
25%	-	(82.11, 83.03)	(81.75, 83.39)	(87.11, 88.46)	(84.52, 85.67)
50%	-	(79.50, 81.06)	(80.44, 81.76)	(84.32, 86.08)	(82.85, 84.97)

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Table 4.28 shows the 95% confidence intervals for the classification tree performance of surrogate splits for MAR and MNAR. As the missing rate increased, the accuracies of all the classification trees for the samples are also decreasing.

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

Percentage of Missing Values	MAR	MNAR
5%	(84.65, 85.91)	(80.25, 82.21)
10%	(84.51, 85.40)	(78.57, 79.74)
15%	(84.57, 85.46)	(80.62, 81.81)
25%	(83.27, 84.23)	(80.37, 81.31)
50%	(82.04, 83.23)	(79.43, 80.76)

Figure 4.32 shows that as the missing rate increased, the accuracies of all the classification trees for the dataset have an obvious trend of decrease. In general, for samples that contain only one predictor variable missing, the accuracy is gradually decreased as the missing rate increased. On the other hand, for samples that contain 2 predictor variables missing, the effect on the accuracies is very small if the percentage of missing data in the dataset is less than 15 percent. When the percentage of missing data is between 15 percent and 50 percent, there is an

obvious decrease in the classification accuracy. However, for samples that contain 3 to 4 predictor variables missing, the effect on the accuracies is very small if the percentage of missing data in the dataset is less than 25 percent.

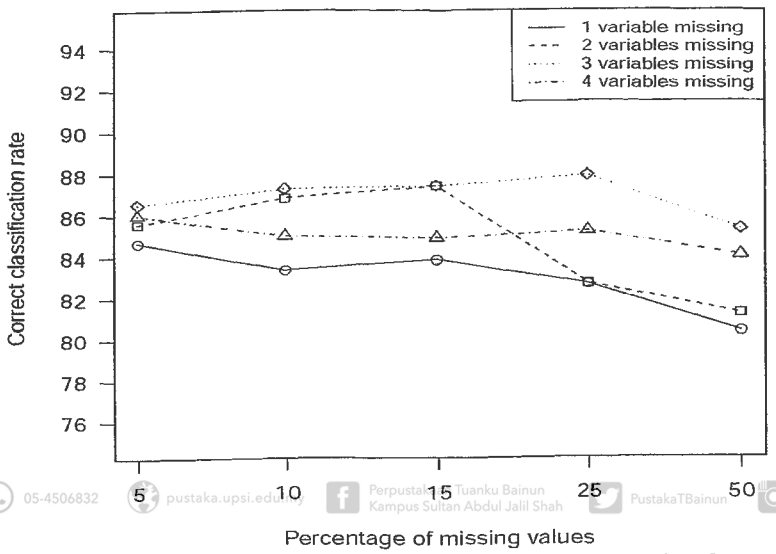


Figure 4.32: Percentage of correct classification rate in dataset with different level of missing values

Figure 4.33 shows that as the missing rate increased, the accuracies of all the classification trees for the dataset have an obvious trend of decrease. As expected, samples that contain MNAR generate classification tree with less accuracy compared to MAR.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

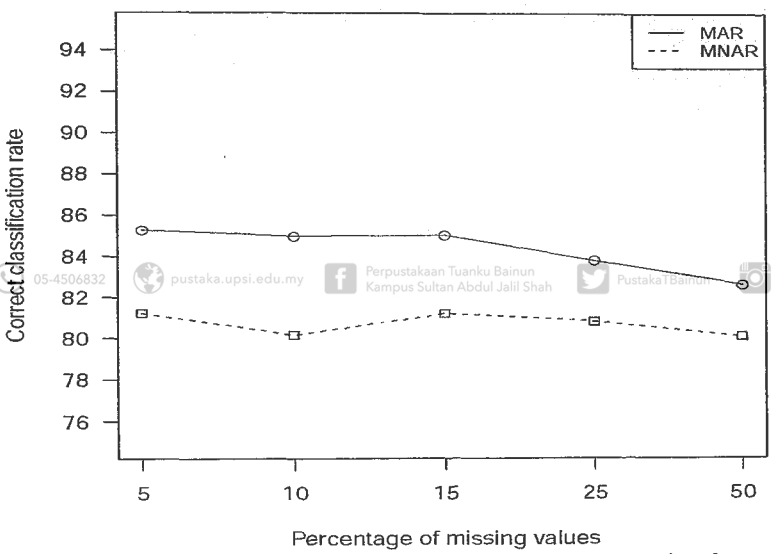


Figure 4.33: Percentage of correct classification rate in dataset with different level of missing values

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

4.12 Conclusion

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

The exploratory data analysis presents a lot of useful information about the university data. The mosaic plots and meta-analysis presents a lot of useful information about the performance between female and male students. The meta-analysis proves that there are gender difference in academic performance for first class degree, second class upper degree and second class lower degree. The result of predicting students'academic performance using classification tree is shown to outperform the bayesian network and naive Bayes.

The results show that classification tree structure is sensitive to the percentage of missing values for this large sample of students' background data. The splitting rules used are changed even for five percent missing values while the tree structure is maintained if the percentage of missing values are equal or less than 10 percent. Despite the fact that the classification tree model is sensitive, the classification accuracy is not much reduced. The predicted value from classification can be used to impute the missing values for this large sample of students' background data. It is shown that for dataset with 15 percent missing values, classification tree with imputation is better than surrogate split in order to obtain more accurate tree.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

CHAPTER 5

MODEL DEVELOPMENT

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

5.1 Introduction

This chapter provides a detailed explaination on model development. Section 5.2 presents the development of classification tree with imputation model. Section 5.2.2 explain the imputation of missing values using three imputation techniques; classification tree, Bayesian Network and naive Bayes. All three imputation techniques will be implemented on datasets having three types of missing values mechanism; MCAR, MAR and MNAR.

5.2 Development of Classification Tree with Imputation Model

The main idea of our proposed model is to incorporate the missing data imputation in constructing a classification tree. Firstly, we divide the dataset into two sub-dataset; sub-dataset having records with missing values, $D_{missing}$ and sub-dataset without missing values, $D_{complete}$. Secondly, we modeled the $D_{complete}$ using three imputation methods namely classification tree, Bayesian network and naive Bayes by choosing the variable of missing value as the leaf class nodes. After that, we fill in the missing values in $D_{missing}$ according to the leaf class nodes produced by the imputation model. The most accurate imputation methods is selected to develop classification tree with imputation model. After imputation, we renamed the $D_{missing}$ data as D_{impute}. D_{impute} is then combined with $D_{complete}$ to construct the classification tree with imputation model. The algorithm to construct classification tree with imputation model is given in Section 5.2.1.

5.2.1 Algorithm for Classification Tree with Imputation Model

The procedure to construct classification tree model with imputation can be described in four steps.

- STEP 1
  Divide the dataset into two sub dataset:
  - Dataset contains missing values, $D_{missing}$
  - Dataset without missing values, $D_{complete}$
- STEP 2
  - Missing value imputation algorithm (Section 5.2.2)
- STEP 3
  Combine $D_{complete}$ and D_{impute}. Renamed as $D_{combine}$
- STEP 4
  Construct the classification tree using $D_{combine}$.

pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

5.2.2 Algorithm for Missing Values Imputation

In this study we consider three types of missing values mechanism; MCAR, MAR and MNAR. For each types of missing values mechanism, we run and compare the imputation performance using three imputation techniques namely classification tree, Bayesian network and naive bayes.

5.2.2.1 Algorithm for Imputation using Classification Tree

Firstly, we identified predictor variable having missing values, X_i, $(1 \leq i \leq M)$ where M is the total number of predictor variables in $D_{missing}$. For each predictor variable X_i we built a tree using CART algorithm for the sub-dataset $D_{complete}$ by setting X_i as the leaf class nodes. For example, if the variable with missing values is FACULTY, we construct classification tree to classify student's FACULTY based on CATEGORY, AGE GROUP, RACE and DEGREE CLASSIFICATION.

Figure 5.1 shows an example of the classification tree used to impute the missing values in variable FACULTY.

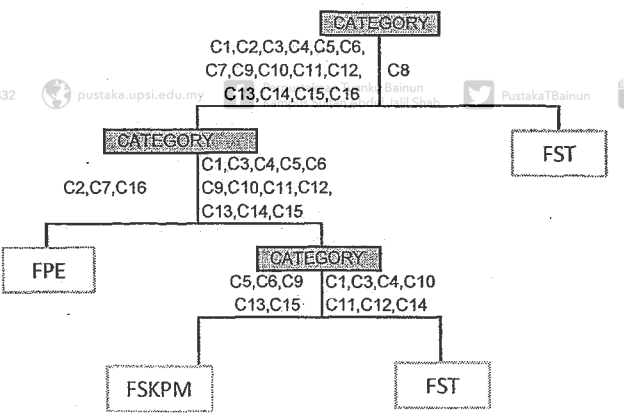


Figure 5.1: Classification tree used to impute the missing value in variable FACULTY

For multivariate imputation cases, we use different trees to impute each of the missing predictor variables. For example, if there are q predictor variables missing, we need to construct q imputation trees. The splitting variables at the top level of the tree are chosen to be imputed first. The terminal nodes class values are used as the predicted value for missing data. Finally, we reconstruct the classification tree using both non missing cases and cases with imputed values. The

standard tree is compared to the tree with imputed values.

Suppose 4 predictor variables in sub-dataset $D_{missing}$ contain missing values, we have 4 different imputation trees. We begin the imputation for variable FACULTY followed by AGE GROUP, RACE and finally CATEGORY. Figures 5.2, 5.3, 5.4 and 5.5 ilustrate the imputation trees used to impute missing values in variables FACULTY, AGE GROUP, RACE and CATEGORY.

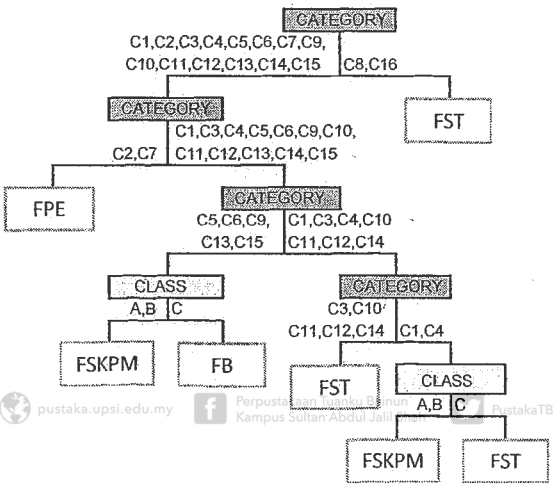


Figure 5.2: Classification Tree to impute missing data in variable FACULTY

The algorithm to construct classification tree for imputation is as follow:

- Input:
  - A dataset $D_{missing}$ with missing values
  - A dataset $D_{complete}$ without missing values
- Output: A dataset D_{impute} with all missing values imputed
- STEP 1
  - Find X_i, $i = 1, 2, ..., M$ where M is the total no. of predictor variable, in $D_{missing}$, having missing values.
  - Build a classification tree T_{impute} from $D_{complete}$ considering X_i as the class predictor.
- STEP 2

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

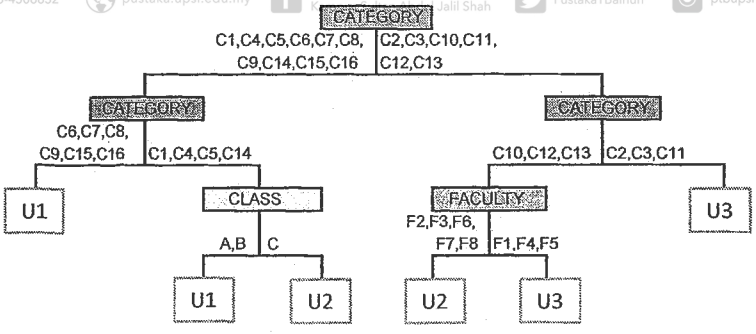


Figure 5.3: Classification Tree to impute missing data in variable AGE GROUP

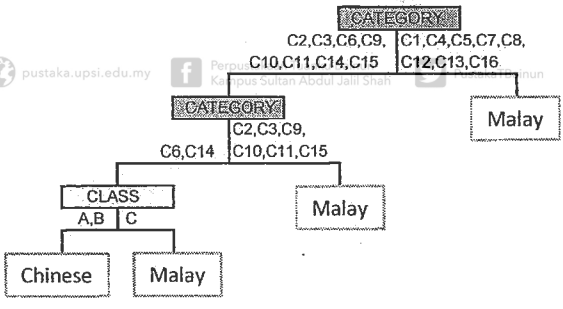


Figure 5.4: Classification Tree to impute missing data in variable RACE

- Impute all the missing values in X_i using the terminal node classification produced by T_{impute}.
- Renamed the dataset having imputed values as D_{impute}.

5.2.2.2 Algorithm for Imputation using Bayesian Network

A Bayesian network for a set of variables $X = X_1, X_2, ..., X_k$ is represented by an acyclic directed graph. Each node in the graph represents a random variable and there is a conditional distribution $P(X_i|X_{pa(i)})$ for each variable X_i given its parents in the graph, $X_{pa(i)}$. A main property of Bayesian networks is that the complete joint distribution of X given as:

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

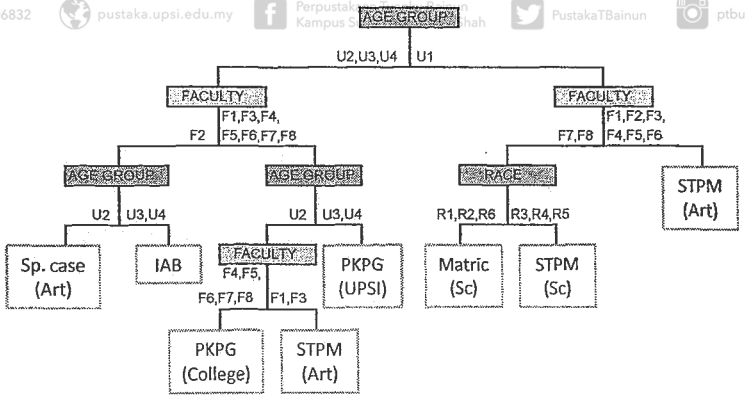


Figure 5.5: Classification Tree to impute missing data in variable CATEGORY

$P(X) = \prod_{i=1}^{k} P(X_i|X_{pa(i)})$

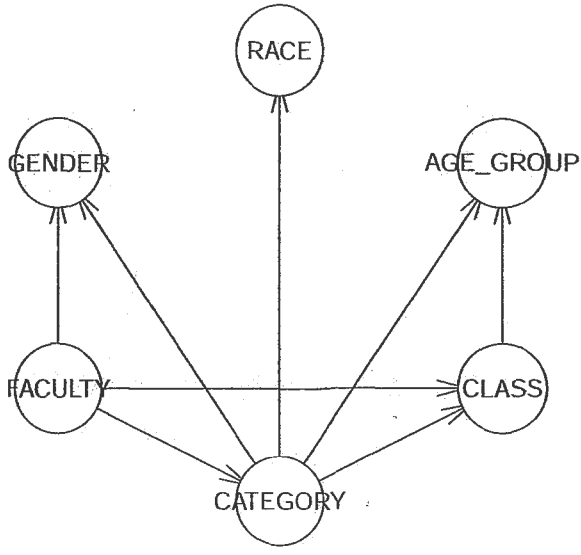
Firstly, the Bayesian network model is built using the complete sub-dataset, $D_{complete}$. After that, the model is used to simulate all the missing values. In general, imputation using Bayesian network can be divided into two steps; Learning a model from the data, and imputing the missing values using the learnt model. In this thesis we applied the K2 algorithm proposed by Romero and Salmern (2004)

The imputation of missing values of a variable is done based on the estimated conditional probability distribution of the variable given its parents. If variable X_i has missing value, we randomly generate one value from the conditional distribution $P(X_i|Pa(X_i))$, where $Pa(X_i)$ represents the observed values of the parents of X_i. For example, if variable FACULTY has missing value, we randomly generate one value from $P(FACULTY)$ because variable FACULTY has no parent node. If variable CATEGORY has missing value, then we randomly generate one value from conditional distribution of CATEGORY given the the value of its parent, that is FACULTY. The example of Bayesian network learnt from $D_{complete}$ is shown in Table 5.6.
The algorithm for imputation using Bayesian network is as follow:

- Input:
  - A dataset $D_{missing}$ with missing values
  - A dataset $D_{complete}$ without missing values
- Output: A dataset D_{impute} with all missing values imputed
- STEP 1:

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi



05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Figure 5.6: Bayesian network learnt from complete sub-dataset

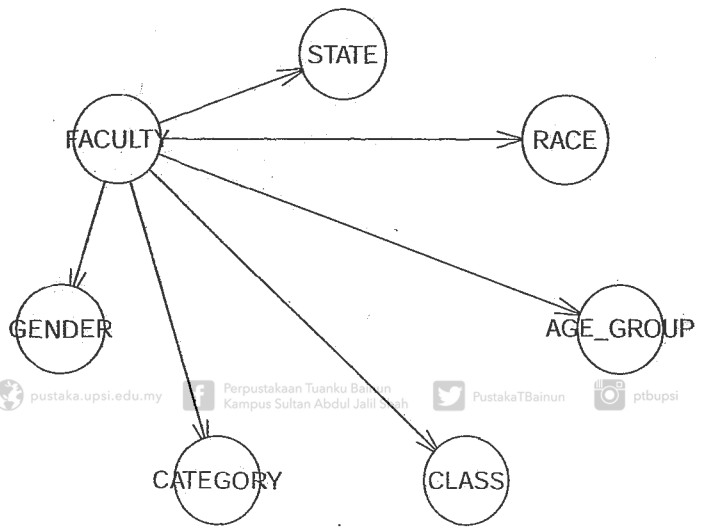
  - Build Bayesian network model using $D_{complete}$

- STEP 2
  - Set M, where M is the variables having missing values
  - For $i = 1$ to M
  - Learn a Bayesian network from the model obtained in STEP 1
  - For each missing values, generate a new value from the distribution $P(X_i|Pa(X_i))$
  - Replace the missing value with the generated value

5.2.2.3 Algorithm for Imputation using Naive Bayes

Naive Bayes is aaplied to predict the most probable value for each missing values in predictor variables given the evidence provided by the observed records of each class. This process is repeated for all predictor variables having missing values. For each predictor variable X_i we built a naive Bayes classifier for the sub-dataset

$D_{complete}$ by setting X_i as the predicted class. For example, if the variable with missing values is FACULTY, we construct naive Bayes classifier to classify student's FACULTY based on CATEGORY, AGE GROUP, RACE and DEGREE CLASSIFICATION.



05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Figure 5.7: Naive Bayes classifier to impute missing data in variable FACULTY

The algorithm for imputation using naive Bayes is as follow:

- Input:
  - A dataset $D_{missing}$ with missing values
  - A dataset $D_{complete}$ without missing values
- Output: A dataset D_{impute} with all missing values imputed

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

- STEP 1
  - Find X_i, $i = 1, 2, ..., M$ where M is the total number of predictor variables in $D_{missing}$, having missing values.

- Build a Bayesian network model NB_{impute} from $D_{complete}$ considering X_i as the class predictor.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

- STEP 2
  - Impute all the missing values in X_i using the Bayesian network model produced by NB_{impute}.
  - Renamed the dataset having imputed values as D_{impute}.

5.3 Conclusion

This chapter presents a detailed explaination on the development of the classification tree with imputation model. We explain the imputation techniques using classification tree, Bayesian network and naive Bayes for three types of missing data mechanism; MCAR, MNR and MNAR.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

CHAPTER 6

EXPERIMENTAL RESULTS

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

6.1 Introduction

This chapter presents the results of experiments applied to real students' background and academic performance dataset to evaluate the performance of proposed algorithms. Sections 6.2, 6.3 and 6.4 present the imputation results of three missing values mechanism; MCAR, MAR and MNAR using classification tree, Bayesian network and naive Bayes. Finally, we present the result of classification tree with imputation model in Section 6.5.

6.2 Result of Imputation using Classification Tree

This section presents the imputation of MCAR, MAR and MNAR using classification tree. All types of missing data mechanism show an increase in accuracy after imputation as shown in Tables 6.1, 6.2 and 6.3.

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classification (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MCAR	No	Yes	Yes	71.86
3	Tree with 15% missing data after imputation	No	No	Yes	73.42

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classi-fication (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MAR	No	Yes	Yes	72.03
3	Tree with 15% missing data after imputation	No	No	Yes	73.12

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classi-fication (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MNAR	No	Yes	Yes	70.24
3	Tree with 15% missing data after imputation	No	Yes	Yes	71.33

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

6.3 Result of Classification Tree with Imputation using Bayesian Network

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

This section presents the imputation of MCAR, MAR and MNAR using Bayesian network. The classification accuracy slightly increase for MCAR and MAR after imputation. However, the accuracy of classification tree reduced for imputation of MNAR using Bayesian network. The result is summarized in Tables 6.4, 6.5 and 6.6.

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classification (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MCAR	No	Yes	Yes	71.86
3	Tree with 15% missing data after imputation	No	No	Yes	72.40

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

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

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classification (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MAR	No	Yes	Yes	72.03
3	Tree with 15% missing data after imputation	No	No	Yes	72.12

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classi-fication (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MNAR	No	Yes	Yes	70.24
3	Tree with 15% missing data after imputation	No	Yes	Yes	69.45

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

6.4 Result of Imputation using Naive Bayes

This section presents the imputation of MCAR, MAR and MNAR using naive Bayes. The classification accuracy slightly increase for MCAR after imputation. However, the accuracy of classification tree reduced for imputation of MAR and MNAR using naive Bayes. The result is summarized in Tables 6.7, 6.8 and 6.9.

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classi-fication (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MCAR	No	Yes	Yes	71.86
3	Tree with 15% missing data after imputation	No	No	Yes	72.13

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classification (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MAR	No	Yes	Yes	72.03
3	Tree with 15% missing data after imputation	No	No	Yes	71.78

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

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

	Tree Model	Predictor variables Change? (compared to standard) tree)	Tree Structure Change? (compared to standard tree)	Nodes Change? (compared to standard tree)	Percentage of Correct Classification (%)
1	Tree with complete data (Standard Tree)	-	-	-	73.88
2	Tree with 15% MNAR	No	Yes	Yes	70.24
3	Tree with 15% missing data after imputation	No	Yes	Yes	68.67

Table 6.9: Summary of the classification tree model before and after imputation for MNAR using naive Bayes

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

6.5 Classification Tree Model with Imputation

Figure 6.1 shows the classification tree model with imputation. This classification tree has similar shape as the standard classification tree. The percentage of correct classification produced by this tree model is 73.42 percent which is much closer to classfication accuracy produced by the standard tree.

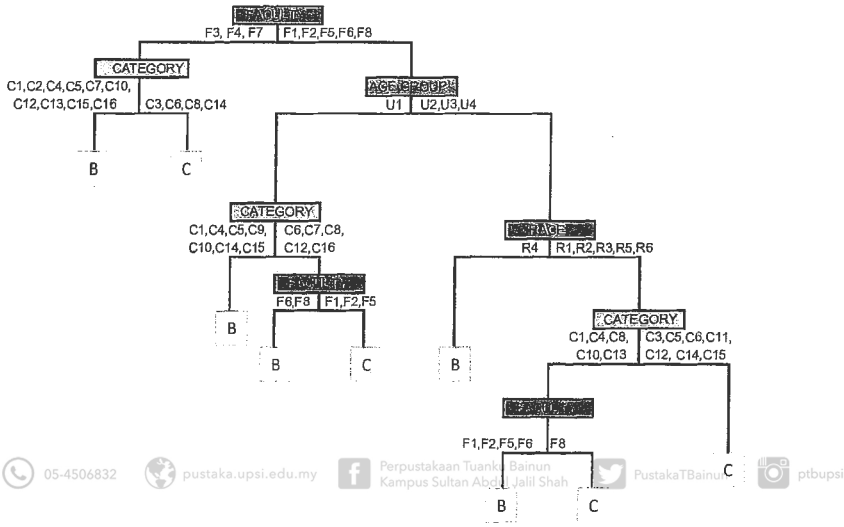


Figure 6.1: Classification tree model with imputation for students degree classification

Figure 6.2 shows the performance of classification tree with imputation. Result proves that imputation using classifcation tree produces the most accurate classification tree compared to imputation using Bayesian network and naive Bayes. Besides that, it is also shows classification tree model with imputation perform better than the standard classification tree using surrogate split when the dataset contains missing values.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Conclusion

The student academic performance is of great concern to the higher educational institution, where several factors may affect the performance. Decision making and planning in the higher educational institution may require extensive analysis of student's performance pattern. Data mining techniques can be used to extract meaningfull information and to develop significant relationships among variables stored in the student's background data.

Statistical analysis is used to see the pattern of academic performance among students from various background. Using cross tabulation, mosaic plot and meta analysis, we found some very interesting results. Female students perform better than male students on average. Female students are more likely to obtain first class and second class upper degree than their male counterparts. The results also reveal that from 1997 to 2008, some faculties such as FPE, FST and FSKPM produce more first class degree students compared to other faculties. Based on the dataset, students from some intake categories and race have not yet received first class degree.

Classification tree has been widely applied in decision making and planning in higher educational institution in recent years. In this thesis we use classification tree to assist the higher educational institution decision making and planning by classifying student's performance using their background data. We show the use of data mining to discover patterns and relationships in student's background data. The result shows that student's background data can be successfully used to construct classification tree model to predict student's performance before they entering the university. It is revealed that faculty, student's intake category, students age and race are the factors that differentiate the performance among students.

The management of higher educational institution need to predict the student's performance precisely in order to obtain a better understanding of student's background and their academic performance pattern. However, in reality, those students records can be missing and in fact, missing values is a common problem. Missing data may affect the classification tree models. We have also shown that nonparametric classification trees algorithm works well on our dataset with at most 10 percent of incomplete observations without requiring data imputation. The result shows that, for all three types of missing data mechanism, the classification tree structure changed and classification accuracy decreases as the missing data increases.

In the study, we proposed the classification tree with imputation model to handle missing values in dataset. We showed that imputation using classification tree produce more accurate result compared to Bayesian network and naive Bayes for

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

all MCAR, MAR and MNAR. The use of multiple trees to impute multivariate missing values is already proved to outperform surrogate split. The application of missing data imputation can help to improve the classification accuracy in classification tree. The imputation tree models used in this study is merely good enough for generating the imputation values and produce better result than surrogate splits.

7.2 Suggestion for Future Research

We have shown that student's background data can be successfully used to construct classification tree model to predict student's performance. As our studies mainly use demographic variables as independent variables, future studies may aim at collecting more variables such as school grade and pre-university grade in improving the academic performance prediction accuracy.

The proportion of each class in the dataset used in this thesis is not balanced since the percentage of first class students is too small compared to other classes. The standard classification tree model will not be able to predict the first class students. We solve this problem by stating the terminal node using probability of each class. Investigating and modifiying the clasification tree algorithm would be a natural extension of this study. Future studies may focus on developing classifiers that are robust and insensitive to skewness in the dataset. Besides that, it would also be interesting to study hybrid algorithms for imbalanced dataset.

The classification tree structure is proved to be sensitive to missing values. Future study may also be extended to study the effect of missing data when missingness occurs in both the training data or testing data or both training and testing data.

In this thesis we show that classfication tree imputation produce better tree in term of structure and accuracy than the surrogate split for 15 percent missing data. Future studies may aim at investigation other imputation methods such as hybrid methods especially for dataset contains more than 15 percent missing data.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Acuna, E. and Rodriguez, C. 2004. The treatment of missing values and its effect in the classifier accuracy. In *Classification, Clustering and Data Mining Applications* (eds. D. Banks, L. House, F. R. McMorris, W. Arabie, and W. Gaul), 639–648. Springer-Verlag Berlin-Heidelberg.

Adeyemo, A. B. and Kuye, G. 2006. Mining Students' Academic Performance Using Decision Tree Algorithms. *Journal of Information Technology Impact* 6 (3): 161–170.

Al-Radaideh, Q. A., Al-Shawakfa, E. M. and Al-Najjar, M. I. 2006. Mining Student Data using Decision Trees. In *The 2006 International Arab Conference on Information Technology (ACIT'2006)*.

Ali, S., Haider, Z., Munir, F., Khan, H. and Ahmed, A. 2013. Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus. *American Journal of Educational Research* 1 (8): 283–289.

Allison, P. 2002. *Missing Data*. Thousand Oaks, California: Sage.

Archer, J., Cantwell, R. and Bourke, S. 1999. Coping at University: An Examination Achievement, Motivation, Self-Regulation, Confidence, and Method of Entry. *Higher Education Research and Development* 18 (1): 31–54.
05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi
Batista, G. E., Prati, R. C. and Monard, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations* 6 (1): 20–29.

Bekele, R. and Menzel, W. 2005. A Bayesian approach to predict performance of a student (BAPPS): A Case with Ethiopian Students. *Algorithms* 22 (23): 24–29.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. *Classification and Regression Trees*. New York: Chapman and Hall.

Brown, H. E. and Burkhardt, R. L. 1999, Predicting Student Success: The Relative Impact of Ethnicity, Income, and Parental Education, AIR 1999 Annual Forum Paper.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Oversampling TEchnique. *Journal of Artificial Intelligence Research* 16: 321–357.

Clayton, G. E. and Cate, T. 2004. Predicting MBA No-Shows and Graduation Success with Discriminate Analysis. *International Advances in Economic Research* 10: 235–243.

Cortez, P. and Silva, A. 2008. Using Data Mining to Predict Secondary School Student Performance. In *Proceedings of 5th Annual Future Business Technology Conference*, Brito, a. and teixeira, j. edn., 5–12. EUROSIS, chapter 9.
Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah ptbupsi

Creel, D. and Krotki, K. 2006. Creating Imputation Classes Using Classification Tree Methodology. In *Proceedings of the Survey Research Methods Section, ASA.*

Cumming, G. 2011. *Understanding The New Statistics: Effect Sizes, Confidence Intervals and Meta-Analysis*. New York: Taylor & Francis Group, LLC.

Delavari, N., Phon-Amnuaisuk, S. and Beikzadeh, M. R. 2008. Data Mining Application in Higher Learning Institutions. *Informatics in Education* 7 (1): 31–54.

Diggle, P. J. and Kenward, M. G. 1994. Informative Dropout in Longitudinal Data Analysis. *Applied Statistics* 43 (1): 49–93.

Ding, C. S., Song, K. and Richardson, L. I. 2006. Do Mathematical Gender Differences Continue? A Longitudinal Study of Gender Difference and Excellence in Mathematics Performance in the U.S. *Educational studies* 40 (3): 279–295.

Ding, Y. and Simonoff, J. S. 2010. An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data. *Journal of Machine Learning Research* 11: 131–170.

Ferrari, P. A. and Barbiero, A. 2012. Simulating ordinal data. *Multivariate Behavioral Research* 47: 566–589.

Gibb, S. J., Fergusson, D. M. and Horwood, L. J. 2008. Gender differences in educational achievement to age 25. *Australian Journal of Education* 52 (1): 63–80.

Glass, G. V. 1976. Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher* 5 (10): 3–8.

Graduate Tracer Study Report 2009. Retrieved 14/11/2012, Website, http://www.mohe.gov.my/portal/en/penerbitan-kpt-selanjutnya/394-tracer-study-report-2009.html.

Graham, L. 1991. Predicting Academic Success of Students in a Master of Business Administration Program. *Educational and Psychological Measurement* 51: 721–727.

Han, J. and Kamber, M. 2001. *Data mining: Concepts and techniques*. London: Morgan Kaufmann.

Hashim, J. 2012. Academic excellence as selection criteria among Malaysian employers. *Higher Education, Skills and Work-based Learning* 2 (1): 63–73.

Hayes, K., King, E. and Richardson, J. 1997. Mature Students in Higher Education: III Approaches to studying in access students. *Studies in Higher Education* 22 (1): 19–31.

Hortan, N. J. and Kleinman, K. P. 2007. Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician* 61: 79–90.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Iacus, S. M. and Porrob, G. 2007. Missing data imputation, matching and other applications of random recursive partitioning. *Computational Statistics & Data Analysis* 52: 773–789.

Ibrahim, Z. and Rusli, D. 2007. Predicting Students' Academic performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression. In *21st Annual SAS Malaysia Forum.*

Ismail, N. A. and Awang, H. 2008. Assessing the Effects of Students' Characteristics and Attitudes on Mathematics Performance. *Problems of Education In The 21st Century* 9: 34–41.

Ismail, N. A. and Othman, A. 2006. Comparing University Academic Performances of HSC Students at the Three Art-Based Faculties. *International Education Journal* 7 (5): 668–675.

Jabor, M. K. A., machtmes, K., Kungu, K., Buntat, Y. and Nordin, M. S. 2011. The Influence of Age and Gender on the Students Achievement in Mathematics. In *International Conference on Social Science and Humanity*, 304–308. IACSIT Press.

Jo, T. and Japkowicz, N. 2004. Class imbalances versus small disjuncts. *SIGKDD Explorations* 6 (1).

Kubat, M. and Matwin, S. 1997. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 179–186. Morgan Kaufmann.

Kumar, S. A. and Vijayalakshmi, M. N. 2011. Efficiency of Decision Trees in Predicting Student's Academic Performance. In *Computer Science & Information Technology (CS & IT) series*, 335–343. Computer Science Conference Proceedings.

Lietz, P. 2006. A Meta-Analysis of Gender Differences in Reading Achievement at the Secondary School Level. *Studies in Educational Evaluation* 32: 317–344.

Lindberg, S. M., Hyde, J. S. and Petersen, J. L. 2010. New Trends in Gender and Mathematics Performance: A Meta-Analysis. *Psychol Bull.* 136 (6): 1123–1135.

Little, R. J. A. and Rubin, D. B. 2002. *Statistical Analysis with Missing Data.* 2nd edn. New York: John Wiley and Sons, Inc.

Longadge, R., Dongre, S. S. and Malik, L. 2013. Class Imbalance Problem in Data Mining: Review. *International Journal of Computer Science and Network (IJCSN)* 2 (1): 2277–5420.

Munnich, R. and Schurle, J. 2003, On the simulation of complex universes in the case of applying the German Microcensus, DACSEIS Research Paper Series No 4, University of Tubingen.

Naderi, H., Abdullah, R., Tengku Aizan, H., Sharir, J. and Kumar, V. 2009. Creativity, Age and Gender as Predictors of Academic Achievement Among Undergraduate Students. *Journal of American Science* 5 (5): 101–112.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Nghe, N. T., Janecek, P. and Haddawy, P. 2007. A Comparative Analysis of Techniques for Predicting Academic Performance. In *37th ASEE/IEEE Frontier in Education Conference*.

O'keefe, D. J. and Haie, S. L. 2001. An Odds-Ratio Based Meta-Analysis of Research on the Door-in-the-Face Influence Strategy. *Communication Reports* 14 (1): 31–38.

Peiperl, M. and Trevelyan, R. 1997. Predictors of Performance at Business School: Demographic Factors and the Contrast Between Individual and Group Outomes. *Journal of Management Development* 16 (5): 354–367.

Peng, L., Lei, L. and Naijun, W. 2005. A Quantitative Study of the Effect of Missing Data in Classifiers. In *Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on*, 28–33. IEEE.

Ried, K. 2006. Interpreting and Understanding Meta-Analysis Graphs A Practical Guide. *Australian Family Physician* 35 (8): 635–638.

Romero, V. and Salmern, A. 2004. *Multivariate Imputation of Qualitative Missing Data Using Bayesian Networks. Advances in Soft Computing*, vol. 26. Springer Berlin Heidelberg.

Rubin, D. 1976. Inference and Missing Data. *Biometrika* 64 (3): 581–592.

Saar-Tsechansky, M. and Provost, F. 2007. Handling Mising Values when Applying Classification Models. *Journal of Machine Learning Research* 8: 1625–1657.

Shafer, J. 2002. *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.

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.

Sterne, J. A. C. and Harbord, M. 2004. Funnel plots in meta-analysis. *The Stata Journal* 4 (2): 127–141.

Stevens, T., Olivarez Jr, A., Lan, W. L. and Tallent-Runnels, M. K. 2004. Role of Mathematics Self-Efficacy in Mathematics Performance Across Ethnicity. *The Journal of Educational Research* 97 (4): 208–221.

Vandamme, J. P., Meskens, N. and Superby, J. F. 2007. Predicting Academic Performance by Data Mining. *Education Economics* 15 (4): 405–419.

Vateekul, P. and Sarinnapakorn, K. 2009. Tree-Based Approach to Missing Data Imputation. In *ICDM Workshops*, 70–75.

Vialardi, C., Bravo, J., Shafti, L. and Ortogisa, A. 2009. Recommendation in Higher Education Using Data Mining Techniques. In *Proceeding of the 2nd International Conference on Educational Data Mining (EDM'09)* (eds. T. Barnes, M. C. Desmarais, C. Romero, and S. Ventura), 190–199. www.educationaldatamining.org.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

APPENDIX A

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Description for the code used in the tree diagram

Variable	Code	Description
Faculty	F1	Faculty of Languages
	F2	Faculty of Business and Economics
	F3	Faculty of Cognitive Science and Human Development
	F4	Faculty of Music
	F5	Faculty of Sports Science
	F6	Faculty of Human Sciences
	F7	Faculty of Science and Technology
	F8	Faculty of Information Technology and Communication
Category	C1	Diploma
	C2	Institut Aminuddin Baki
	C3	KDPK
	C4	Special/Diploma (Science)
	C5	Special Case/Diploma (Arts)
	C6	Overseas
	C7	Matrikulation (Accounts)
	C8	Matrikulation (Science)
	C9	MPSAH
	C10	PKPG (College)
	C11	PKPG (UPSI)
	C12	Second Channel
	C13	Teacher Certificate
	C14	STPM (Science)
	C15	STPM (Arts)
	C16	Change University

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

Variable	Code	Description
Race	R1	Bumiputra Sabah
	R2	Bumiputra Sarawak
	R3	Chinese
	R4	Indian
	R5	Malay
	R6	Origin Native
Age Group	U1	29 or below
	U2	30 - 39(U2)
	U3	40 - 49(U3)
	U4	50 and above(U4)
Degree Classification	A	First Class
	B	Second Class Upper Division
	C	Second Class Lower Division

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi