WINSORIZE TREE ALGORITHM FOR HANDLING OUTLIERS IN CLASSIFICATION PROBLEM

CH'NG CHEE KEONG

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2016

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

CH'NG CHEE KEONG

calon untuk Ijazah PhD
(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"WINSORIZE TREE ALGORITHM FOR HANDLING OUTLIERS IN CLASSIFICATION PROBLEM"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : *03 Ogos 2015.*
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
August 03, 2015.

Pengerusi Viva: *(Chairman for VIVA)*	Prof. Dr. Zurni Omar	Tandatangan *(Signature)*	
Pemeriksa Luar: *(External Examiner)*	Assoc. Prof. Dr. Mohd Rizam Abu Bakar	Tandatangan *(Signature)*	
Pemeriksa Luar: *(External Examiner)*	Dr. Wan Rosmanira Ismail	Tandatangan *(Signature)*	
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	Dr. Nor Idayu Mahat	Tandatangan *(Signature)*	

Tarikh:
(Date) *August 03, 2015*

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Pepohon pengelasan dan regresi (CART) direkabentuk untuk meramal atau mengelas objek dalam kelas yang telah ditentukan daripada suatu set pembolehubah peramal. Namun, kewujudan unsur pencilan mampu menjejaskan struktur CART, ketulenan dan ketepatan peramalan dalam pengelasan. Sebahagian penyelidik memilih melakukan kaedah pra-pencantasan atau pasca-pencantasan pada CART untuk mengendali unsur pencilan. Kajian ini mencadangkan algoritma pepohon pengelasan terpinda, dikenali sebagai pepohon *Winsorize* berdasarkan taburan kelas dalam set data latihan. Pepohon *Winsorize* menyiasat semua unsur pencilan yang mungkin dalam data dari nod ke nod sebelum memeriksa titik pembelahan untuk mendapatkan nod dengan ketulenan tertinggi. Batas atas dan batas bawah plot kotak telah digunakan untuk mengenal pasti unsur pencilan dengan nilai ekstrem melebihi Q ± (1.5×Julat antara kuartil). Data pencilan yang telah dikenalpasti akan dineutralkan menggunakan kaedah *Winsorize* manakala indeks Gini *Winsorize* kemudian digunakan untuk menghitung kecapahan dalam kalangan taburan kebarangkalian bagi nilai peramal yang disasarkan sehingga kriteria henti ditemukan. Kajian ini menggunakan tiga petua henti: nod yang telah mencapai 10% minimum daripada jumlah set latihan, n_{min}, nod yang mengandungi 70% atau lebih kehomogenan dan indeks Gini *Winsorize* terhitung antara dan di antara pembolehubah adalah 70% atau lebih. Keputusan yang diperolehi daripada tujuh (7) set data sebenar menunjukkan bahawa pepohon *Winsorize* merekodkan kadar ralat yang sama atau lebih rendah berbanding pepohon tradisional dan pepohon tercantas dalam semua kes terutamanya yang melibatkan pembolehubah yang banyak. Kaedah ini menawarkan proses pengelasan yang lebih baik dengan menyiasat dan mengendali unsur pencilan dalam semua nod. Justeru, sebarang proses pencantasan akan dihentikan apabila kriteria henti dipatuhi. Pepohon *Winsorize* menghasilkan struktur pepohon paling ringkas dan menggunakan bilangan pembolehubah yang sedikit dengan kadar ralat yang rendah. Pepohon *Winsorize* menawarkan sokongan untuk melaksanakan pengelasan kepada pengamal baru dan pengamal berpengalaman mungkin mendapati kaedah ini memudahkan tugas pra pemprosesan dan analisis.

Kata Kunci: Pepohon pengelasan, Data pencilan, Indeks Gini Winsorize, Algoritma pepohon Winsorize

Abstract

Classification and Regression Tree (CART) is designed to predict or classify the objects in the predetermined classes from a set of predictors. However, having outliers could affect the structures of CART, purity and predictive accuracy in classification. Some researchers opt to perform pre-pruning or post-pruning of the CART in handling the outliers. This study proposes a modified classification tree algorithm called Winsorize tree based on the distribution of classes in the training dataset. The Winsorize tree investigates all possible outliers from node to node before checking the potential splitting point to gain the node with the highest purity of the nodes. The upper fence and lower fence of a boxplot are used to detect potential outliers whose values exceeding the tail of Q ± (1.5×Interquartile range). The identified outliers are neutralized using the Winsorize method whilst the Winsorize Gini index is then used to compute the divergences among probability distributions of the target predictor's values until stopping criteria are met. This study uses three stopping rules: node achieved the minimum 10% of total training set, n_{min}, node contains 70% or above of homogeneity, and the computed Winsorize Gini purity index within and between variables is equal or greater than 70%. The results obtained from seven (7) real dataset indicate that the Winsorize tree scores equal or lower error rates than the traditional and pruned trees in all cases especially when dealing with many variables. This method offers a better classification process by investigating and handling the outliers in all nodes. Therefore, it does not require any pruning process as it stops once the stopping criteria is met. The Winsorize tree produces the simplest tree structure and it typically uses fewer variables with a low error rate. It offers some assistance for performing classification to new practitioners and experienced practitioners may find this method simplify preprocessing and analysis tasks.

Keywords: Classification tree, Outliers, Winsorize Gini index, Winsorize tree algorithm

Acknowledgement

I would like to express my sincere appreciation to Dr Nor Idayu binti Mahat for her valuable effort, guidance, patience, support and encouragement in supervising this work. Warms thanks to Prof. Madya Dr Sharipah Soaad Syed Yahaya and Dr Nazrina Aziz for providing valuable information regarding to statistics.

I would like to thanks the various people in School of Quantitative Science, Universiti Utara Malaysia as they provided me a very useful and helpful assistance.

Special thanks to the librarians who are always willing to lend their hands to get my requested books and articles. Thanks to Ch'ng Li Guat for rescuing me in computer problems.

I am grateful to all my friends for cheering me up the working room and thanks to them for the friendship, caring and entertainment.

This study would not have been possible without financial support. I would like to thank the JPA and SLAI which has supported me during my study. Also, thanks are addressed to Universiti Utara Malaysia for giving me the opportunities in completing my works.

The appreciation also goes to my parents, Ch'ng Seow Khin and Tan Pheik Sim, my family members and my wife, Low Joon Khim for their emotional supports, love, motivation and caring during my study. This thesis is dedicated to them. Thanks in million again to all for providing me a loving environment.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ARM	Association Rule Mining
CART	Classification and Regression Tree
CHAID	Chi-Square Automatic Interaction Detection
EBP	Error Based Pruning
HER	Electronic Health Record
ESD	Extreme Studentised Deviate
HMM	Hidden Markov Model
ID3	Iterative Dichotomiser 3
ILPD	Indians Liver Patient Dataset
IQR	Inter Quartile Range
L	Lower Boundary
LOF	Local Outlier Factor
LOFB-DRF	Extension of Random Forest
LS	Least Square Method
MAD	Median Absolute Deviation
MD	Mahalanobis Distance
MDL	Minimum Descriptive Length
MED	Median
MEP	Minimum Error Pruning
ML	Maximum Likelihood
MR	Map Reduce
OR	Outlier Region
SP	Splitting Point
U	Upper Boundary

List of Appendices

CHAPTER ONE
INTRODUCTION

1.1 Introduction to Classification

Classification is a scientific process that refers to activities of allocating objects into pre-determined classes. Also, it is attempting to identify to which group or class a new object should belong to. The classification can be distinguished into two types: *unsupervised classification* and *supervised classification* (Gupta, 2006). Unsupervised classification refers to the process of defining classes of objects where one usually aims at either identifying some explainable structures among objects or looking for convenient partitions of the collection of objects. Unlike supervised classification, there are no explicit target attribute which associated with the input. Two examples of simple classical statistics method of unsupervised classification are clustering and dimensionality reduction (Ghahramani, 2004). Often, the number of hypothesized number of clusters ahead of time will be set by the users (Duda, Hart & Stork, 2001).

In contrary, supervised classification is the process of allocating a new object into its predefined class. The concept of supervised classification is as follows: a classification rule that decides to which class an object should be assigned will be constructed based on a set of measurements obtained from the classified objects (Cunningham, Cord & Delany, 2008). Then, the constructed classification rule will be evaluated in order to ensure that it is suitable to classify (or to predict) the class of a future object. The interest of supervised classification is to search for the best possible algorithm that will be able to produce a general hypothesis to predict the correct class of the future objects (Kotsiantis, 2007; Chaovalit & Zhao, 2005). The challenge in supervised classification is to build a concise and accurate mathematical model that can assign future object into a correct class. They are many type of supervised classification methods such as decision tree, support vector machines, logistic discrimination, naïve Bayes, random forest, neural network, ensembles, perceptron and much more.

This thesis is concerned with the supervised classification thus the discussion throughout this thesis will refer classification as supervised classification.

1.2 Examples of Classification Problem

In business activities, classification approach can be used to explore the behaviour of buyers (brand loyalty) and to determine market segmentation (Miller, 2005, p. 25 & 26). One may need to understand the buyers' behaviour towards a certain product or brand. Some criteria including salary, marital status, gender and types of occupation may be used to explain one's preference either interested or not interested to purchase a new brand of product. In banking sector, based on the customer profile, the industry is using a particular classifier to evaluate the risk of approving loan (Thomas, Oliver & Hand, 2005). The well-known classification algorithms were used to investigate the credit score data sets accurately (Baesens, Gestel, Viaena, Stepanova, Suykens & Vanthlenen, 2003).

Besides, classification is widely used in medicine. In hospital for example, a doctor is assisted by a systematic rule to claim a survival rate of heart plant patients (Gupta, 2006, p.15). Also, it has been used to classify human chromosomes into its respective groups (Curnow & Franklin, 1973). Different types of classifier have been applied to detect anomaly intrusion in system. The classifier can overcome security threats in computer network and can be used to identify unauthorized use of computer system (Bahrololum & Khaleghi, 2008). In other areas, decision tree (C4.5) technique is used in stock management and control (Wu, Lin & Lin, 2006). With the growth of online information, Joachims (2005) used Support Vector Machine (SVM) for text categorisation where the main goal is to classify documents into fixed number of predefined categories. Meanwhile, K-nearest neighbours algorithm is studied to diagnose on the Wiscousin-Madison breast cancer (Sarkar & Leong, 2000).

1.3 Classification Rules

Many classification rules have been devoted by researchers such as Fisher discriminant function, decision trees, neural networks, nearest neighbour approaches, logistic discriminant, and naïve Bayes classification. Each rule has its strengths in dealing with various structures of the data, among others include distributions of population i.e. population's distribution, type of variables and correlation among variables. Despite of variety classification rules, the oldest systematic classification procedure called decision tree has become a focus of interest in this thesis. Decision tree is a logical model which often represented as a binary tree (two-way split). It shows how the target variable can be predicted using all independent variables. The tree straightforward shows how each independent variable is split, which then lead to the prediction of the target variable. Such interesting feature has made this tool unique compared to other existing classification tools which commonly explained by mathematical formulae. Like continuous development on most classification tools, this thesis concentrates to investigate the decision tree, often termed as tree, in an attempt to add some values in the methodology of constructing it.

1.3.1 Elements of Decision Tree

In classification problem, the goal of a decision tree is to predict the value of a target variable, which represented group of objects using some input variables. Figure 1.1 shows a simple decision tree with two splits. A basic structure of a tree includes (i) *node* and (ii) *branch*. A tree begins with a *parent node* (labelled "`a`") and it splits into two *non-terminal nodes* (labelled "`b`" and "`c`"). The binary split from the previous node is called "branch". For example, parent node "`a`" produces a branch that contains node "`b`" and node "`c`". Each non terminal node will split continuously until it cannot be split due to some predetermined constraints. The node that can be split is termed as non-terminal node whilst the final node which cannot be split anymore is called *terminal node* or *leave*, i.e. nodes with labelled "`e`", "`f`", "`g`", "`h`" and "`i`".

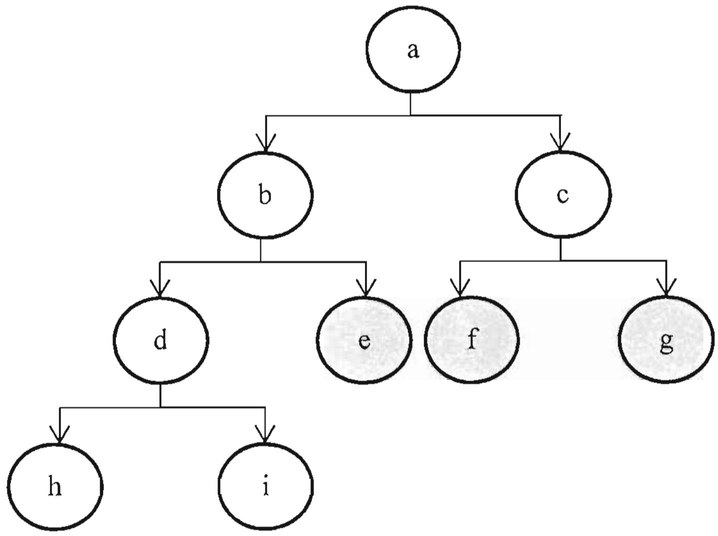


Figure 1.1. Simple decision tree

The structure of tree is simple but produces a powerful form of multiple of variable analysis. It is a flow chart like structure which split from node into branch like segment by its algorithm. There are many types of tree available in practices including ID3 (Iterative Dichomotomiser 3), CHAID (Chi-squared Automatic Interaction Detection) and CART (Classification and Regression Tree). The difference about these trees lay on criteria used in the splitting process. ID3 is a tree based on information theory and attempt to minimize the expected number of comparisons. The first question asked must divide the search into two large domains while the subsequent perform a little division of the space (Dunham, 2003). However, ID3 has many disadvantages where it can only deal with nominal variables, unable to deal with noisy data as it could lead to overfitting tree structure, incapable to handle missing values, always end up with bushy tree and much more. (see Xu, Wang & Chen, 2006; Octavian, 2011). Therefore, C4.5 was devoted to improve the condition of ID3. Later, another type of tree called Chi-square Automatic Interaction Detector

(CHAID) was popularised by Kass in year 1980. It is built for non-binary tree which is used for large dataset.

In comparison to ID3 and C4.5, CHAID performs *Chi-square* test and *F*-test for classification and prediction purposes. CHAID is normally used in direct marketing (Haughton & Oulabi, 1997) and it is said a perfect tool to discover the relationship between variable (Gilbert, 2010). Another structure of tree called classification and regression tree (CART) has interesting features where it only performs binary split in every single split of tree construction. Such structure supports high speed deployment and considered by many as the most versatile predictive modelling algorithm which produce an accurate prediction. Besides, it may consider various types of variables in a single structure hence makes it as a good choice of tree in many real practices (Loh, 2011; Breimen, Friedman, Olshen, & Stone, 1984). Therefore, this study sets to focus more on this type of tree.

In general, CART is much simpler than CHAID and C4.5 as it does not split into multi-ways. Moreover, it will keep on splitting until the specified threshold is met. The splitting process of CHAID might be stopped too early as this method attempts to avoid over fitting. Thus, some of unimportant variables might be masked by important variables. Meanwhile in C4.5, the pruned tree will be just substituted by a branch which caused insufficient of information. And, all errors are treated as equal which in practical application, some errors are might be more serious than the others. Although decision trees are much easier to be understood, the process of constructing a good, accurate and reliable decision tree is influenced by the data.

1.3.2 Construction of Decision Tree

Decision tree can be learned by splitting data into subset based on the attribute value test. A set of if-then rules is used to improve the human readability. The tree like graph is used for inductive inference. It is said to be robust to noisy data and capable of learning disjunctive expression. It also provides a highly effective structure where it could balance the risks and rewards associated with each possible course of action. The data is split randomly into training set and test set where the former set is used to construct a tree and the latter is used to evaluate the constructed tree. The use of training set and test set will avoid the construction of over-performed tree and will provide a reliable tree for future classification. The tree is built in accordance with a splitting rule which divide the data into smaller part where the objects from the same class are assigned into the same nodes. This process is repeated on each derived subset by top-down induction of decision tree until each leaf consists of a single observation (Rokach & Maimon, 2008), and this scenario is referred as maximum homogeneity (Breimen et al, 1984). *Gini index*, *Entropy*, and *Twoing* splitting rules are commonly used as a splitting algorithm to separate the objects in every node. Among these algorithms, *Gini index* is widely used as it works well for noisy data especially in classification tree. This index is computed for each variable and the one with the highest *Gini purity index* (or lowest Gini impurity) will be selected for the next variable to be split.

Specifically, a tree begins with a parent node, t_p. The parent node t_p will be spilt into left (t_l) and right (t_r) child nodes by using the best splitting value of variable, x_j^R. The process of splitting is repeated at both left and right child nodes to produce more child nodes. Such processes are repeated until either a tree or every node reaches a pre-determined threshold. The maximum tree means only one class in the terminal node. However, the maximum tree may turn out to a very huge, complicated and bushy which may have hundred levels. Thus, setting a threshold is needed. In this case, the splitting is stopped when the number of objects in the node is less than a predefined required minimum, n_{min}. Usually, n_{min} is set as 10% (Timofeev, 2004) of the learning sample size. To get the maximum right size of tree, pruning procedure is applied which consider on the optimal proportion between the complexity of tree and the error rate.

1.4 Classification and Regression Tree (CART)

Classification and regression tree (CART) is among the popular classification methods which proposed by Breimen et al. (1984). This type of tree tackles two types of variable where a classification tree is suitable for categorical dependent variable and regression is suitable when dependent variable is quantitative (Wilkinson, 1992). CART algorithm uses a multistage decision process by completing a set of variables jointly to make a decision. On top, there is a root of tree or called parent node that would split into binary ways (0 for left split and 1 for right split) which associate with the internal nodes (child nodes). Decision would be made based on the threshold at every level. As depicted in Figure 1.2, objects in the parent node (t_p) will be split into either the right internal node (t_r) or the left internal node (t_l). Let x_j be the splitting value of the variable X_j at t_p. The split occurs such that objects in the t_r will have values of X_j greater than the splitting value, x_j, and objects in the t_l will have values of X_j equal or smaller than x_j.

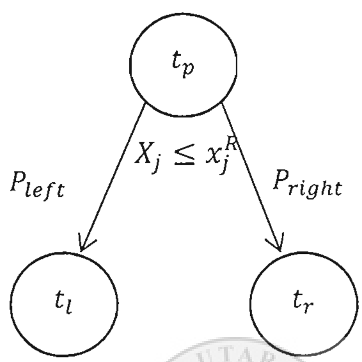


Figure 1.2. Splitting algorithm of CART

Bertolini (2006) demonstrated how a classification tree can be used as an effective tool for quality control practices in oil pipelines. The tree identifies which pipelines to monitor and to choose the most suitable monitoring policies for it. His study was motivated by Breiman et al. (1984) which suggested that inspection activities and spillage can be detected or recognised by operating classification and regression trees method. The idea provides a better way to detect the expected spill for cross country oil pipelines although different countries face different types of failures. Breimen et al. (1984) indicated that digit recognition can also be done by using classification and regression tree (CART). Besides, CART has been used to characterize the long-term survival after surgery (Valera, Walter, Yokohama, Koyama, Liai & Okamoto, 2006). Chen, Wang and Zhang (2011) used tree in biometric and statistical genetics. CART also applied in medical diagnosis and prognosis. Breimen et al. (1984) used the methods to diagnose heart attacks problem. He tried to classify patients into two different classes: patients who are at risk of dying within 30 days following heart attack (class 1) and the survivor (class 2). CART is ideally suited for exploring and modelling the complexity in ecology data. De'ath and Fabricius (2000) studied on the ecology data sets using soft coral survey data from Australian central Great Barrier Reef. They analyze three groups of taxa which are *Efflatounaria, Sinularia* and *Sinularia Flecibilis.* CARTs have been used to analyse the relationship and partition the response into homogeneous group. Furthermore, CART is also widely used in galaxy classification, financial crisis or defaults, classifying mammals, and so much more.

1.5 Challenges in Constructing a Classification Tree (CART)

In real practices, there is no specific formula to confirm on how good the constructed classification rule is. As the matter of fact, the choice of "good classification rule" depends on the perspective, background, intuition or intention of practitioners in constructing the rule (Jacobs, 2001). Some practitioners aim to have a rule that will give minimum cost of loses rather than a rule with the highest accuracy of classifying objects to their correct group. Some would strongly rely on the accuracy indicators (e.g. error rate and Brier score) where the rule with the highest accuracy is the best choice. Statisticians would evaluate the goodness of classification model based on the mean square error and variance of estimator. Sometime, the priority of choosing a rule is based on the simplest one and much easier to be understood. Statisticians, economists and medical practitioners put much effort to work with a linear base-rule due to its straight forward process whilst machine learning groups and engineers would prefer on rules that do not rely on any standard assumption such as normality of data. Therefore, there is no exactly the best rule but the process of classification technically searches for the best possible rule.

Outliers are extreme data points which have the potential to influence the statistical analysis (Evan, 1999; Jacobs, 2001). The occurrence of outliers may due to mistake made during data entry or in fact valid. Simply ignoring the outliers would destabilise the estimation. Therefore, the whole data must be routinely inspected so that the true colour of the outlier can be defined accurately. Although there are many analytical calculation and graphical displayed tools to spot the outliers, some type of outliers might be masked by several reasons. How if we do not minimise the distortion? And, what would happen to the quality of the data if no action to be taken to such outliers? How the tree structure would be if the data contains outlier? Unreliable output will be generated from the unfiltered data. Outlier may bring a huge effect to some rule's construction. For instance, a slightly different value in the data would create a different tree classifier. Figures 1.3 to 1.6 are the examples of Kyphosis and Iris data sets to demonstrate how the construction of trees can be deviated due to the influence of outliers. Ignoring such outlier problem may result in wrong estimated values hence producing different structure of trees. At worst, a future object may be allocated to an incorrect class.

In the Kyphosis data set, three variables are used to classify objects into two levels of kyphosis (a type of deformation) either absent or present after the operation. The constructed trees based on data without outliers (Figure 1.3) and with outliers (Figure 1.4) indicate that different trees have been constructed due to the influence of outliers though the same variables have been chosen in the tree construction. Sometimes, the existing of outliers may influence the choice of variable to be split. In the example of Iris data set as shown in Figure 1.5 and Figure 1.6, the outliers reflect the changes on the parent node. The examples given give a sign that somehow outliers may influence the structure of the constructed tree.

The possible challenge in this problem is that the object might be misclassified into a wrong group. Figure 1.3 and Figure 1.4 demonstrate the classification process on Kyphosis data set. The data have three independent variables (`Age`, `Start`, `Number`) and a dependent variable (type of deformation) with two states, absent or present after the operation. Figure 1.3 shows the constructed tree without outliers while Figure 1.4 shows the tree with outliers. Both trees have different structures on the left split due to the present of outliers. Although the outlier is small, it may give some impacts on the structure of the tree, the splitting points and future classification. If a future object has criteria with `Age` = 36, `Start` = 11, `Number` = 3, then tree in Figure 1.4 will assign such object to group of present but tree as in Figure 1.3 will identify it as absent. For this reason there is a need to properly address the occurrence of outliers in a tree.

Commonly, measuring the accuracy of a constructed tree can be done by taking the error rate and the cost of error. But, the latter is sometime hard to achieve as prior information or expertise knowledge is required.

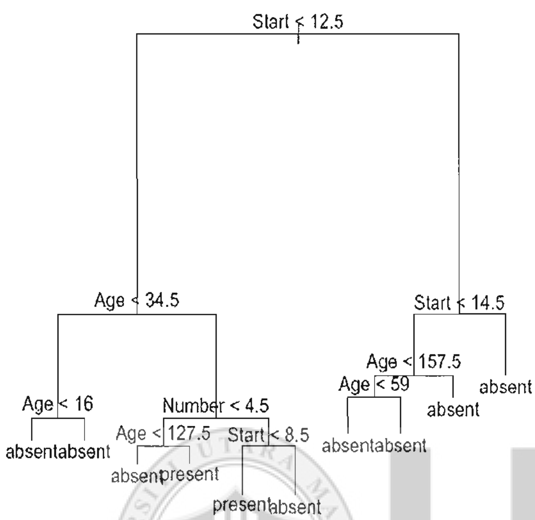


Figure 1.3. Tree classifier for Kyphosis (without outlier)

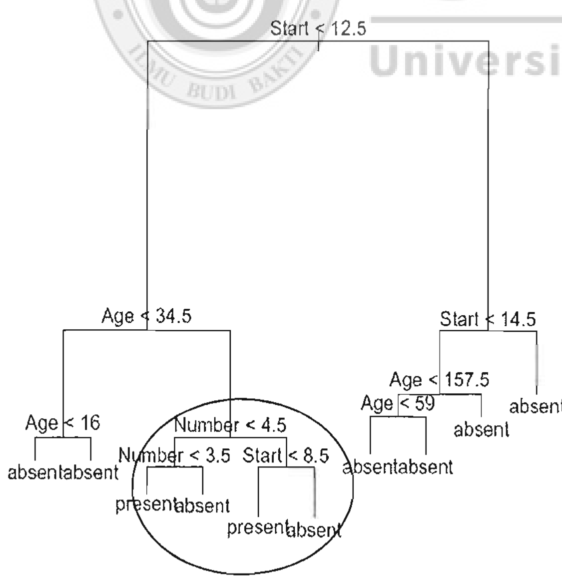


Figure 1.4. Tree classifier for Kyphosis (with outlier)

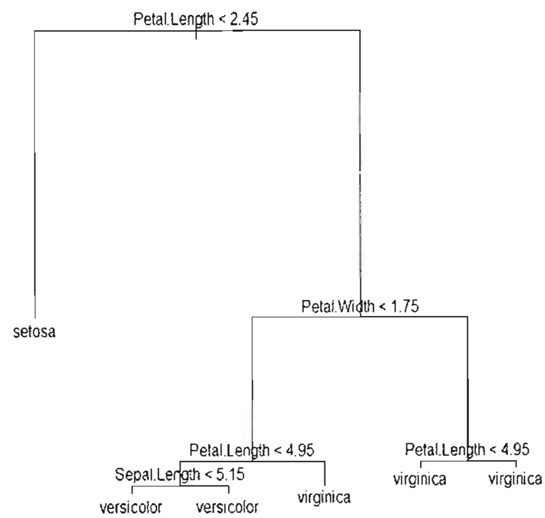


Figure 1.5. Tree classifier for Iris (without outlier)

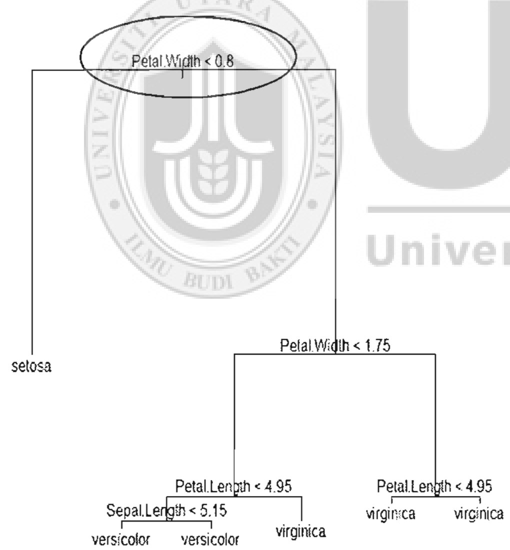


Figure 1.6. Tree classifier for Iris (with outlier)

1.6 Problem Statement

Sections 1.3 and 1.4 have given a general idea about CART and some challenges on constructing such type of tree have been highlighted in Section 1.5. Unawareness towards the existence of outliers can cause a misguidance to the future cases as the constructed model is bias and inaccurate to present the behaviour of actual information (Tabia & Benferhat, 2008). Thus, many researches have been carried out to solve the problem in classification. The first approach is to depend on the tree itself to isolate the outliers from process of splitting the data during tree construction. This approach was implemented by Breimen et al. (1984) and Shouman, Turner and Stocker (2011). Then, the tree is pruned accordingly to reduce the complexity of the tree classier and hence improves the predictive accuracy by the reduction of overfitting. However, using all the data may lead to a bias and bushy decision tree (John, 1995; Engels & Theusinger, 1998). Thus, pruning the constructed tree would be a good option. Although pruning process could produce an accurate tree with balance size of tree, but this method requires pruning knowledge and experience. It demands users with some statistical or analytical knowledge, but could be troublesome to practitioners. Therefore, some researchers prefer to perform a pre-processing before constructing a tree (Reif, Goldstein, Stahl & Breuel, 2008; Kyung, June, Dao & Nam, 2011; Han & Kamber, 2006). In pre-processing phase, graphical tools such as Boxplot and probability plot would be used to identify outliers in the data. Once the outliers are determined, then the next step is to critically handle them. Eliminating the outliers is easy but it produces "clean data set" which will definitely provide us a "good classifier". As pictured in Figure 1.3 to Figure 1.6, few outliers can cause a tremendous bias split to the whole structure of tree classifier. Sometimes, this phenomenon will be even worse when some of the explanatory variables are masked by the outliers. It means that the "extreme value" might hide some variables to be split. Therefore, this method could be a risk if the tree contains bias split as wrong classifier might provide wrong prediction to us. John (1995) has introduced an idea of pruning and reconstructing tree. The branches of tree will be pruned at the first time to eliminate the outliers. Then, the tree will be reconstructed in order to get the fitted tree. Despite of promising and unbiased tree, the idea faces with some drawbacks as it demands for double tasks to prune and reconstruct the tree. Besides, many data could be lost due to outliers' termination. Wang, Gu and Wang (2004) suggested another idea that developing a tree by starting with the most insensible attribute (the attribute that give the less important in classification). As the tree growth, the most sensible (most important attribute) will be chosen hence the outlier will be isolated in some nodes at the bottom of the tree. Yet, this idea has received little attention from other researchers

Considering the weaknesses of earlier idea or approaches, this study initiates the idea of reducing the effects of outliers using a method called Winsorize, which commonly used to compute robust statistics e.g. mean, standard deviation and etc. The idea of winsorizing is to set all the outliers to a specified percentile of the data. However, the choice of percentile is subjective. Too low percentile will allow the outliers to be included in the tree construction but too high percentile will lead to small variance of measurements but high bias to the tree. The idea of when to accommodate outliers is another issue to debate. If one performs Winsorize on the data prior to construction of a tree, then we will miss out to see the state of outliers in a tree. Such phenomenon happen because the outliers have been replaced with the percentile before the classification is taken place. To allow a tree that represents the actual data, this study proposes to have simultaneous processes of detecting and winsorizing outlier as well as nodes splitting. We winsorize the data when the outlier is found so that the splitting algorithm namely *Gini index* can be computed without the influenced of the detected outliers. Then, we split the original data using the estimated *Winsorize Gini Purity Index.* This proposed strategy will promise a splitting process that is not biased towards skewed data which lead to produce full unbiased structure of tree. This structure will explain about the data and will be useful for future data especially when the future data also contain outliers.

Selecting the right variable and the splitting point are important in order to get a maximum homogeneity in every single split. The maximum homogeneity of left and right child node from previous node is equivalent to the change of impurity function, $\Delta i(t)$. It means that the objects which have the similar behaviour are assigned into their own group. However, the outliers would just affect the purity and cause to a bias structure of tree at the end. Therefore, the process of constructing a tree that is not sensitive towards outliers needs to be outlined. This study is looking for the best possibility to the tree structure.

Generally, tree is allowed to split as bushy as it could in order to achieve maximum homogeneity. Then, the tree is pruned based on the tolerant error rate. However, this could lead to time consuming. Alternative to this practice, this study suggests to stop the splitting process before over fitting tree is obtained.

1.7 Research Objectives

This study proposes a new algorithm of tree that insensitive towards the outliers. Therefore, the research objectives of this study are:

1. To determine outlier in a data prior to construct the branch of tree.
2. To manage the identified outliers accordingly using Winsorize method.
3. To integrate the process of determining outlier and identifying outliers with the recursive process of constructing a tree
4. To propose stopping criteria in constructing tree in order to avoid an over-fitting tree.
5. To compare the new Winsorize tree with the traditional trees.

1.8 Significant of Study

This study provides an alternative classification rule based on decision tree suitable to handle the contaminated data. It offers a data cleaning process embedded in the classification process, which is better than common practices that clean the data prior to the classification. Such simultaneously processes may highlight outliers in the classification, identified by the simple information extracted from a box plot at each investigated node. Next, the proposed Winsorize Gini purity index offers an unbiased way to deal with selection of information variable for splitting. Whilst, the stopping criteria suggested in this study may assist on constructing a tree at optimum level without waste.

In practice, sometimes a practitioner may have some doubts with the data in hand especially when outliers are detected. Some outliers occur due to mistake in data entry, measurement error or in fact valid. Simply ignoring or terminating the suspected values could be a risk which might cause violation to the end result. Therefore, this study provides a process which insensitive towards outliers making the computed error rate less biased. Overall, the proposed tree construction strategy ensures a quality data used for data mining, which will be helpful for practitioners or researchers whom are less proficient with tree methods.

1.9 Scope of Study

This study focuses on the problem of constructing decision tree for classifying objects into one of two groups when a sample is contaminated with outliers. Current practices need practitioners to clean the data before a construction of tree. Such practice demands a practitioner to master the arts for cleaning the data to avoid over-cleaning which may end up with over performance in classification. Besides, the choice of tool for identifying outliers may not comply with the aim of classification, to minimize the error for future data. In fact, some practitioners might be too relying on the tree itself as it could isolate the outliers into separated nodes. However, this scenario might end up with a bushy tree and some important variables might be masked. This study aims on improving such practices by performing the process of data cleaning and construction of a rule simultaneously to offer much convenience and reliable used among practitioners. However, detection of outliers was set among the continuous variables rather than categorical variables. The continuous data is sorted and the suspected value according to the preceding and succeeding values is then examined. The detected outlier will be penalised before performing the Gini measurement for splitting. Although there are various types of trees, this study uses the CART which performs as binary split. This tree could perform classification with multi-type of variables, thus make it as a convenience tree for practices.

1.10 Thesis Organization

This thesis focuses upon the problem of the outliers while constructing the tree to obtain a more reliable and accurate tree. This chapter describes the background of tree and highlights the problem facing in the method when dealing with outliers. Also, this chapter mentions about the contribution towards the body of knowledge in both academic and industrial.

Chapter two of this thesis reveals the parametric and nonparametric models in classification. It draws the attention on why the previous classification tree method is not performed well when dealing with outliers in the data. Also, the chapter discusses some outliers detection and handling methods which have been widely used since few decades ago. Besides, the research gaps, the benefits and drawbacks of trees are also illustrated in details in this chapter.

The foundation of the proposed method is displayed in chapter three where it examines the previous works and improving in the algorithms and arithmetic in Winsorize Gini index measurement, which contribute to more accurate and precise result in both classification and prediction. These will be the base for the contribution of this study. Besides, the data descriptions are also presented.

Chapter four shows all the results collected from the designed tree on some existing data sets, using the designed research methodology. Comparison between traditional trees, traditional pruned tree and the proposed tree were performed to give evidence that the proposed tree is comparable, and sometimes better than the established tree designs.

The last chapter gives the summary of the study, contributions, limitations, recommendations and possible future works. The successful accomplishments of research objectives are also explained.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This chapter overviews some existing classification rules and the outlier identifiers, the strengths and weaknesses of each method are highlighted.

2.2 Classification Rule

The general term of classification is a process of assigning the objects into their group or category. Before the technical specific methods for classification, people classified the object based on their intuition. Those having the same behaviour or characteristics would be assigned into the same group. However, the intuitive decision would create a serious problem as different people have different intuition. Classification rules have been successfully implemented to solve the real world problem (Mahat, 2006).

Freitas (2014) indicated that classification normally uses prediction rules to express knowledge. IF-THEN rules are used in prediction rules with the condition to produce y, a label given to class or group. If all the condition in antecedent rule are satisfied then the prediction of the goal attribute will be satisfied the consequent rules. With a few conjunctions of if-then rules, the relation between the attributes can be narrowing down. This knowledge is useful and intuitively comprehensive for most users.

2.3 Parametric Approaches

Generally, parametric base classifiers are powerful statistical methods that enable to produce an accurate and precise estimation providing that normality assumptions are satisfied. In contrast, nonparametric methods do not require any normality assumption for parameter estimation.

The parametric test often refers to classical or standard test that makes assumptions about the parameter of the population from the selected samples. Some of the parametric approaches include:

2.3.1 Naïve Bayes Method

Bayes method is a key technology that has been used for classification purposes after it was proposed by Bayes (1702-1761). Bayes approach to statistics attempts to fully utilise the available information in order to reduce the uncertainty so that a better decision can be made. The uncertainty means unknown outcomes of various situations. The expression of "it is probable", "the chances are" and so on are always used to deal with the uncertainty condition. When such expressions are quantified, it means one is dealing with "probabilities". Let $P(A)$ and $P(B)$ refer to the probability that event A will occur and event B will occur. $P(A|B)$ is the conditional case which refers to the probability A would happen given that B has already happened. Then, the Bayes theorem is

$$P(A|B) = P(B|A)P(A)/P(B) \tag{2.1}$$

where

$P(A|B)$ = the probability of the object B belonging to class A.

$P(B|A)$ = the probability of obtaining the attribute values B if we know that it belongs to class A.

$P(A)$ = the probability of any object belongs to class A without any other information.

$P(B)$ = the probability of obtaining the attribute values B whatever class the object belong to.

This method is not sensitive to irrelevant variable, it can handle real and discrete data, more accurate as prior class probability is used and handles stream data well. However, this method has been criticised as it requires us to specify a prior distribution for all the unknown parameters. In many cases, the prior knowledge is vague, unclear, or non existent thus making it extremely hard to specify a value for the model (Duda & Hart, 1973).

2.3.2 Regression

Regression is a statistical method used to describe the nature of relationship between independent variables and a dependent variable. The relationship can be positive or negative, linear or nonlinear. Whilst, correlation is used to determine the relationship between the two variables (x, y) (Bluman, 2004, p. 495; Larson & Farber, 2006, p. 458; Abraham & Ledolter, 2006). A positive relationship means that either variables increase or decrease at the same time whereas a negative relationship means one variable increases but the other decreases and vice versa (Bluman, 2004). The simple linear regression consists of only one independent variable corresponds to one dependent variable. In the multi-linear regression, there is only one dependent variable but several in independent variables.

The equation of linear regression can be written as

i. Simple linear regression

$$y = \beta_1 x + \beta_0. \tag{2.2}$$

ii. Multiple linear regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_0. \tag{2.3}$$

This method can help us to predict the value of one unknown variable through one or more predetermined variable(s). When the relationship between the independent variables and dependent variable are linear, it shows an optimal result. However, linear regression is often inappropriate for non linear relationship. Besides that, the output is only limited to numeric value. The implementation of regression for classification can be done by discretising the numeric dependent variable such that values lower than a threshold belong to class 1, and the remainings to class 2. However, such exercise will be troublesome is the classification involves more than two classes. Further discussion relating to this idea can refer to (Groβ, 2003, p. 33; Seber, 1977; Bluman, 2004, p. 495; Larson & Farber, 2006, p. 458).

2.3.3 Logistic Regression

Logistic regression is another parametric approach that resembles linear regression. In multiple logistic regressions, it describes the relationship between one dependent variable and several independent variables (covariate). What distinguishes a logistic regression from linear regression is that the output is in binary or dichotomous. Individuals whose predicted value probability is more than 0.5 will be assigned to group 1; otherwise to another group. The assumption here is each observation, y_i comes from Bernoulli distribution with $E(y) = P(y = 1)$. The specified form of logistic regression model can be written as

$$p(y = 1) = \frac{\exp(x'\beta)}{1+\exp(x'\beta)} + \varepsilon. \qquad (2.4)$$

Logistic regression has several advantages over the linear regression in classification. For instance, normal distribution assumption is not required in independent variables. It does not assume linear relationship between independent variables and the dependent variable. Besides that, independent variables can be in mixed variables. Unfortunately, in order to get a meaningful and stable result, it needs more data and it might be costly. Other work can be obtained in Hosmer and Lemeshow (2000).

2.3.4 Linear Discriminant Analysis

Linear discriminant analysis was devised by Fisher in year 1936 with the main idea of finding projection to a line which the samples from different classes can be well separated. It also seeks to reduce the dimensionality. Consider assigning an object with measurement vectors x consisting p variables to either class G_1or G_2. A function

$f(\boldsymbol{x})$ of the measurements is used to compare with the threshold to decide which class of the object is classifying to, G_1 if $f(\boldsymbol{x})$ is greater than the threshold and to G_2 otherwise.

Seeking a scalar y by projecting the sample $\boldsymbol{x}$ onto a line $\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x}$. Of all the possible line from each point to the line, select the one with the maximum separability. Measurement of the separability is needed to find the good projection vector. If the means of $\boldsymbol{x}$ in G_1 and G_2 are $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ then the mean of y in G_1 and G_2 can be written as $\boldsymbol{w}^T\boldsymbol{\mu_1}$ and $\boldsymbol{w}^T\boldsymbol{\mu_2}$ respectively. Assuming the covariance matrix, Σ from both group are the same then the variance of $\boldsymbol{Y}$ is $\boldsymbol{w^T\Sigma w}$ in both group and the maximum $\boldsymbol{w}$ is

$$\phi(\boldsymbol{w}) = \frac{(w^T\mu_1 - w^T\mu_2)^2}{w^T\Sigma w}. \quad (2.5)$$

The parameters $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ and $\boldsymbol{\Sigma}$ are usually unknown, thus the estimated parameters are used to replace it. For instance, $\boldsymbol{\mu_1}$ is replaced by $\bar{x}_\iota$ and $\boldsymbol{\Sigma}$ is replaced by $\boldsymbol{S}$ the estimated pooled-covariance matrix. Then the distance measure between two groups is

$$D = \frac{(\boldsymbol{w^T}\bar{x}_1 - \boldsymbol{w^T}\bar{x}_2)^2}{\boldsymbol{w^T}\mathbf{S}\boldsymbol{w}}. \quad (2.6)$$

The best value of w is to choose the maximize $D(w)$ which is given by $\boldsymbol{S^{-1}}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$. There, $\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x}$ can be written as $\boldsymbol{y} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\boldsymbol{S}^{-1}\boldsymbol{x}$. Allocation of an object to G_1 if $\boldsymbol{y}$ is closer to $\bar{\boldsymbol{y}}_1 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_1$ and to G_2 with $\bar{\boldsymbol{y}}_2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}_2$

otherwise. Further discussion on this topic can be found in Lachenbrunch (1975) and Mclachlan (2004).

2.3.5 Advantages and Disadvantages of Parametric Approaches

Generally, parametric modelling has been widely applied to solve real world problems. It is based on the probability distribution which normal distribution is the most common. And, the samples from different groups are independent and the variances are equal between groups. If all the assumptions are satisfactorily then parametric methods produce high accuracy of estimation. The training sites are reusable and it generates information classes. Besides, parametric test is also more powerful than non parametric test when dealing with continuous variables.

However, this approach contains some drawbacks. Parametric is not strong enough when particular assumptions are not met or violated. In addition, we need to consider the cost and difficulties of selecting the training site and the signature homogeneity of information classes might also varies. Moreover, it is only reasonable to apply parametric approach if the sample size is large enough; otherwise nonparametric approach is recommended for those situations. In fact, in real life, the distribution of the data is normally unknown and it is almost impossible to get the data which is normal distributed. Christina (2009) commended that data is often non-normal in the biomedical sciences because the sample size is normally small and the data is having heavy tail, skewness, multimodality, and extreme asymmetries. She recommended that that non-parametric test where it is more appropriate due to the assumptions free tests. Altman and Bland (2009) insisted that the importance of t-test diminishes when the sample size increases. According to Bridge and Sawilowsky (1999), to evaluate the medical literature effectively, statistical test play an important roles on research outcomes. However, applying inefficient statistics is not only increases the need for resources, but more importantly increases the probability of committing a Type I or Type II error. In medical field, t-test is considered as the most prevalent tests used under the normal curve theory. However, parametric test could be violated especially when the assumptions of normality is not met. They suggested non parametric test such as Wilcoxon Rank-Sum test to the violation from population normality. Similar to classification, implementing parametric classification rules for small data or when the true populations' distributions are unknown would be misleading.

2.4 Nonparametric Approaches

Nonparametric approach allows a relaxation of assumption which means it does not rely on any assumption, parameterised distribution and parametric estimation. We use it when the parameter of the variable of interest in the population is unknown. There are several nonparametric methods have been widely used in current studies.

2.4.1 Neural Network

According Lisboa (1992), neural network is an artificial technique which attempts to simulate the neural system. It mimics to the human brain where the neurons are linked together via dendrites. Dendrites, hillock zone, axon, cell body and synapses constituted a biological neuron. Impulses are transmitted through a strand of fiber called axon. Analogous to human brain, an artificial technique (Artificial Neural Network) has been created to solve the classification technique. The artificial neuron is a simple mathematical model which consists of input nodes and output nodes. (Schurmann, 1996). For multilayer artificial neural network, several hidden layers (black box) are embedded between input nodes and output nodes. The network will use some different types of function such as Sigmoid function, Tanh function, Sign function and Linear function. The weighted links are used to strengthen the connection between neurons. It continues to grow in the field of business, scientific, medical and academic world. Neural network requires less formal statistical training, has the ability to detect complex nonlinear relationship between the dependents and independents variables. In addition, neural network can be used for both supervised and unsupervised learning. It also works well with the huge datasets which consist of noisy input data. Sigmoid function makes the data input smooth by handling the anomalies, random error and outlier. Neural network has been widely use in recent research. For instance, neural network was used for survival analysis for personal data. Thus credit scoring pertains to bad or good creditor can be distinguished (Bensen et al, 1995). However, neural network causes greater computational burden and proneness to over fitting (Tu, 1996). Besides that, neural network is lacking the ability to explain its behaviour.

2.4.2 Decision Tree

Decision tree is among the popular classification methods. The output resembles a flow chart like tree structure (Gupta, 2006). It is designed to assist the decision makers to make decision for possible future events. A subsequent decision may occur to encourage the decision maker to think beyond the immediate decision (Coles and Rowley, 1995).

Ho (2004) insisted that decision tree is one of the most useful tools in classification problems. This predictive model constructs a very powerful model in a view of 'tree'. Decision tree consists of a chain of questions. Through the answers to the questions, the accurate goals can be clearly discovered via splitting the complex data precisely into levels. The root or parent node is on top. It splits into branches and creates the child notes until the bottom of node called leaf or terminal node. Decision tree receives much attention because it is easy to generate understandable rule. Decision tree algorithms have been proposed in statistical, machine learning and pattern recognition. Yet, more and more refinements have been implemented to achieve a higher accuracy.

Decision tree provides an easy understanding and interpreting condition. It can handle data which contains mixture type of variables such as continuous and categorical variables simultaneously. It presents the data directly and the tree inherent structure is based on a procedure which can distinguish the useful and useless variables. However, decision trees have some drawbacks such as correlation between the attributes are ignored, tree replications, disable to handle the continuous data accurately and complicacy of bushy trees. Furthermore, the final classification can be deceptive which would lead to misinterpretation. Some variables are not split causing the truth to be masked by other variables (Breiman et al., 1984).

2.4.3 Advantages and Disadvantages of Nonparametric Approaches

The rapid growth of nonparametric statistical procedures over the past six decades was due to its advantages. Hollander and Wolfe (1999) disclose that nonparametric methods forgo the traditional assumptions as in parametric methods (population must be in normal distribution). Besides that, nonparametric techniques are often easier to understand and apply in most of the situations. Furthermore, these techniques are relatively insensitive to outlying observations. Further discussion can refer to Hollander and Wolfe (1999).

Although nonparametric statistical methods contain lots of desirable properties, it seems lacking of power compared to the traditional method as the statistical validation is quite loose. Moreover, currently the appropriate software for nonparametric method is limited. According to Levine (1991) and Simon (1991) in Wilkinson (1992), different commercial programs are always produce different output with the same data. Some programs even worse by providing no documentation and supporting material to explain the algorithm.

2.5 Evaluating Rules

All methods aim to produce a good rule for classification. Many researchers may choose any method to construct a classification rule which is most suitable to their problem. However, which method is the best and reliable? In fact, there is no perfect rule for evaluating the performance. Usually, the evaluation of the performance is done once the classifier has been constructed.

Hand (1997) discussed the evaluating rules used in pratices which include (i) inaccuracy (ii) imprecision (iii) inseparability and (iv) resemblance. Inaccuracy measures the ineffectiveness of the rules in allocating object into the correct groups, while imprecision provides the information between the estimated probabilities $\hat{f}(\pi_i|x)$ and$f(\pi_i|x)$. Inseparability measures how similar are the $f(\pi_i|x)$ belongs to each group at x, average over x. Gini index, Entropy and Chenoff measure are commonly used in inseparability measure where Inaccuracy is equal to the summation of imprecision and inseparability. Finally, resemblance measures the differences between the true probabilities conditioned on the estimated ones. All these aspects have their own strengths, none of them is better criterion than the others but it depends on the aims of the study.

In this research, we are focusing on the inaccuracy measurement due to its simplicity in computation and easy to interpret. Besides, it is the most commonly used indicator by researchers in classification problem.

Suppose the learning set, L, and a sample with n objects, $n \in N$ where each object represented by r $(r = 1,2,3, \dots, n)$, let g_r be the group where the object r comes from $g_r \in \{1,2,3, \dots G\}$ and x_r is the vector of measurements of object r where $x_r \in X$. The learning set is denoted by $L = \{(x_1, g_1), \dots, (x_n, g_n)\}$. The basic idea of inaccuracy is to compare between the original groups of an object and the observed group. The classification rule is considered good if the inaccuracy rate is small. Inverse to inaccuracy is accuracy where the higher rate of accuracy means better classification rule.

The most popular measurement under this umbrella is called misclassification rate or error rate. It calculates the proportions of objects that are misclassified from the classification exercise. Although error rate has few drawbacks such as the cost associated with different kinds of error does not taken into account and the error rate does not penalise the large errors, it is the most popular indicator for evaluating the performance of completing rules (Wang and Johnson, n.d). Types of error rate are discussed in the following sub-section.

2.5.1 Types of Error Rate

Let x be the measurement vectors and g be the class. Let $f(x)$ is the overall distribution of measurement vector x. Let $f(g|x)$ be the probability that a case with measurement vector x will belong to class g.

2.5.1.1 Bayes Error Rate (e_B)

This error rate aims to obtain minimum error rate given a set of measurements. However, this type of error rate can only be obtained if $f(g|x)$ and the posterior probability, $f(x)$ are known. In other words, e_B provides a lower bound on any possible error rate that may be achieved by a real classification rule.

$e_B = \int [1 - max_i f(g|x)] f(x) dx.$ (2.7)

2.5.1.2 Achievable Error Rate (e_b)

Researchers usually use a classification rule without actually knowing the performance of the rule, even the appropriateness of using the rule. They might try to use a rule which they think the best in classifying objects. The error rate computed from the rule is called achievable error rate, e_b which is greater than e_B in general.

2.5.1.3 Conditional Error Rate (e_c) and Unconditional Error Rate (e_E)

These two types of error rates define the sample-based classification rule. Let region R_g is the rules of allocating objects to G and let $\hat{f}(x)$ and $\hat{f}(g|x)$ be estimated from the training set with the assumption of $(R_1 \cup R_2) \in \boldsymbol{R}$. The conditional rate is specified as

$e_C = \sum_{g R_g}^{G} [1 - \hat{f}(g|x)] \hat{f}(x) dx.$ (2.8)

The unconditional error rate also called actual error rate is the conditional on the training set which is used for classification rule.

Unconditional error rate, e_E is the expectation of the conditional error rate over all the design sets of the same size from the population. It is more suitable to be used before seeing the training set (Mahat, 2006).

$e_E = E(\sum_{g_{R_g}}^{G} [1 - \hat{f}(g|x)] \hat{f}(x) dx).$ (2.9)

Among these error rates, conditional error rate is the most popular type which has been widely used by researchers. More information about e_C will be discussed in the next section.

2.6 Estimating Conditional Error Rate

Breimen et al. (1984) pointed out that the used of same data set for both rules construction and evaluation leads to bias results. Therefore, splitting the data set into training sets and test sets is best to overcome such problem. Hold-out validation is a common method where the observations are chosen randomly from the data set to form the training and test set. There are many possibilities on splitting the data but normally less than one third (no specific theoretical justification has been clarified) of the data is used for validation purposes (Breimen, 1984, p. 11). According to Webb (1999), there are two main purposes of splitting the data. First, the classifier is trained by the training set and is used to provide the estimation of its performance. Second, both training set and test set are used in classifier design. The assessment of the model will be done through the percentage of the error rate estimation. Random sub-sampling method is another method that resembles the hold-out method except that it does not rely on a single test set (Gupta, 2006). The estimation is repeated for several times then mean is computed to get an accuracy of estimation. However, those methods are only suitable for a huge data set.

Cross validation or rotation estimation is an old method which was pioneered by Geisser (1975). There are some types of cross validation to handle the smaller data sets. (Goutte, 1997).

2.6.1 *K*-fold Cross Validation

A sample is divided into K subsamples (or sometimes called folders) where each subsample contains approximately equal proportion. One of the subsamples from K will be taken out in turn as a test set and the remaining $K-1$ subsamples are used as training set to construct a classifier. Then, the constructed classifier is assessed by the test set and the error rate is computed. This process is repeated K times until each subsample have been taken out.

2.6.2 Leave One Out Cross Validation

It is similar to the K fold cross validation but only an object is taken out as a test set while the remaining $n-1$ are treated as training set. It has the advantage of constructing a classification rule using a sample as big as the original one which lead to less bias. Unfortunately, the loops of n times give greater variance to the estimate.

2.6.3 Validation Set

Data set is divided into three sets which are training set, test set and validation set. A validation set is commonly used for estimating parameter in learning algorithms. The best accuracy of the value will be used as the final parameter values.

2.6.4 Jackknife

Jackknife is a method introduced by Quenoullie (1949) to estimate the bias of an estimator. This method resembles leave-one-out method as it also involves the process of omitting each subset in turn. The remaining subsets are used to build the rules. However, this method is used to reduce the bias of estimator hence evaluating the variance of the estimator. Some statistic of interest is computed in each sub set of the data. The average of this subset statistics is compared to the statistic computed from the entire sample in order to estimate the bias of the latter.

Let estimate θ using appropriate algorithm for instance maximum likelihood (ML) or least square method (LS) to obtain an estimate $\hat{\theta}$. Observation from the data is deleted and recalculates the estimate for θ from the remaining $n-1$. $\hat{\theta}_{-i}$ denotes the estimate. The pseudo value is given by

$$S_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}. \tag{2.10}$$

The process has to be repeated for all the observations. The jackknife estimate for θ is the mean of the pseudo values,

$$\tilde{\theta} = \frac{1}{n}\sum_{i-1}^{n} S_i = n\hat{\theta} - \frac{n-1}{n}\sum_{i-1}^{n}\hat{\theta}_{-i}. \tag{2.11}$$

2.6.5 Bootstrap

Efron (1983) conceded that bootstrapping performs better than cross validation. Kardi Teknomo (2006) and Chernick (2008) explained that bootstrapping is sampling with replacement from a sample. Bootstrapping is sampling within the sample. This method is analyzing subsample from the data instead of using subsets of the data like cross validation. The sample is picked randomly from the data set. The selected number is then replaced again into the data and has the same chances to be chosen again. Ultimately, all the selected numbers are used to construct the classifier while the unselected samples are used as the test set. Bootstrapping method is not only for estimating generalization error; it also provides confidence bounds estimation for network output (Efron & Tibshirani, 1993). The .632+ bootstrap is currently popular in performing the estimation of generalization error even though there is a severe overfitting. However, this method can run into problem when $n < p$ where n is the sample size and p is the features or variables. The .632+ bootstrap is quite biased when the sample size is small (Molinaro, Simon & Pfeiffer, 2005). Thus, adjusted bootstrap method has been built to solve this problem. The robustness of this method across the situation provides a least bias comparing to leave-one-out bootstrap and the .632+ bootstrap. (Jiang Wenyu & Simon, 2007).

Each of the discussed procedures for estimating the conditional error rate has its own advantages. The advancement of computer has assisted bootstrap and jackknife procedures, but considering these procedures in this study will require excessive computation time. Similarly, the leave-one-out demands for great computation time with big variance. Therefore, this study chooses the hold out validation ($\frac{2}{3}$ of training set and $\frac{1}{3}$ of test set) to evaluation the error rate, following the suggestion of Breiman (1984).

2.7 Pre-processing

Nowadays, the number of machine learning applications is increasing. Therefore, pre-processing stage seems vital to constitute an obligatory step before constructing a model. This step for sure brings a solution to knowledge discovery in databases problem (Engels, 1996; Engels & Theusinger, 1998). In fact, pre-processing is data cleansing, altering the dimensionality of the data and altering the data quantity (Engels & Theusinger, 1998). This study is focusing on data cleansing process which related to treatment of outliers. Therefore, the following section will discuss about several techniques of outliers detection and outliers handling.

2.8 Outliers

Outliers often refer to the value that is beyond bounds or distributions which are inevitable and drastically effects on data analysis (Young, Valero-Mora & Friendly, 2006). In strict term, outliers are the observation which have a substantially difference from what it supposed to be (Hair et al., 1992). The data that appear surprisingly far away from the main group has been concerned as "unpresentative", "rogue", "spurious", "maverick" or "outlying" observation (Barnett, 1978). Hawkins (1980) defines an outlier as an observation that is distinguish further from other observations and arouse suspicious that it could be generated by different mechanism.

The issues of outliers have been discussed widely since it is unnoticed and invisible in real data, but the advance of computer process may discover some erratic behaviour with these contaminated data. Simply ignoring contaminated outlier can lead to inaccurate estimation (Chambers, Hentges & Qiang, 2004; Gentleman and Wilk, 1975; Rousseeuw & Leroy, 2003) and at worst such distortion can produce unreliable output and the cost of handling the bad data can be enormous (De Veaux & Hand, 2005).

However, the outliers' value must be investigated further since they can be due to data entry error or in fact valid (Chambers et al., 2004). Iglewicz and Hoaglin (1993) mentioned that outliers can be caused by several reasons. Some possible sources are gross recording, incorrect distributional assumption, data contain more structure and unusual observation. Sometime, outliers provide useful information which can help us to improve the quality of the data gathering process and to identify an appropriate model for statistical inferences. Some applications attempt to measure the abnormal behavior (outliers) which apart from the norm (Bolton & Hand, 2002). For instance, credit card and telecommunication fraud can be detected through the suspicious or unusual behaviour in the record. In recent year, hacker will try different ways to penetrate the computer system. Unauthorized value in the data can be used to discover the computer attack or intrusion (Bahrololum & Khaleghi, 2008). Koufakou et al. (2008), outlier can provide information of patients who exhibit abnormal symptoms due to their specific disease or ailment.

Johnson (1998) insisted that no statistician or statistical technique can accurately tell the experimenter what to do with the outliers. Own expert opinion can well inform how to deal with the outlier. The inappropriate representation or the errors may be discounted or even eliminated from the analysis (Hair, Anderson, Tathem & Black, 1992; Johnson, 1998). However, simply deleting or removing the peculiar data can result bias outcome. The investigation of Bessel and Baeuer (1838) that discussed by Barnett (1978) claimed that outliers are nature and should not be rejected. Barnett (1978) indicated that rejection or retention should base on the intention or aim, and how the distortion could influence the analysis. Evans (1999) asserted that we should explore reasons why some of the respondents behaved atypically. Those who behave dishonestly but responded honestly must be included in the data set whereas individuals admit intentionally provide dishonest responses should be deleted from further analysis. Pre-modification of the data by changing the substantial data can also seriously destabilize the estimation. The model created by the "clean data" will definitely provide an "overconfident" classifier which might lead to high significant error. In classification, it is not only referring to the extreme value, it also concerns if a point of a class is misclassify in the middle of another class.

2.8.1 Outliers Detection

Outliers in univariate data has been investigated extensively by many researchers however the term "outlier" would never have the precise and exact definition (Barnett & Lewis, 1984).

Iglewicz and Hoaglin (1993) distinguished three issues in outlier which are outlier labelling, outlier accommodation and outlier detection. Outlier labelling means the potential outlier in the data is flagged for further investigation whereas outlier accommodation refers to the use of statistical techniques which will not be unduly affected by outliers. And, outlier detection is the formal test on the outliers.

Ben Gal (2005) described those outliers detection can be divided into two fields which are univariate method and multivariate method. Univariate is proposed in the earlier works whereas multivariate is mostly used in current body of research. The taxonomy fundamental of outlier detection are parametric method and non parametric method.

Statistical parametric method can be applied for a known underlying distribution or statistical estimate unknown distribution. Those value deviates from the model assumption are assumed as outlier. The drawbacks of parametric method are that it is not suitable for high dimensional data sets or the data sets which the prior knowledge of the data distribution is unknown. Non parametric method is a distance based method which is based on the measurement of local distance. Clustering technique is also used to detect outlier which a small of cluster can be considered as outliers (Kaufman & Rousseeuw, 1990; Ng & Han, 1994; Acuna & Rodriguez, 2004). Non parametric can deal with huge data set and is reliable when the distribution of the data is unknown. And, it also does not rely on assumption of the distributions.

For normality assumption, normal probability plot can be applied. The lower and upper tails of the plot can be a useful graphical technique to identify potential outliers.

Also the plot such as boxplot, stem and leaf and histogram can help us to determine whether it is single outlier or multiple outliers.

Among the existing mathematical formulation in identifying outliers, one of the easiest ways to identify outliers can be done using the boxplot. The main ingredients for the boxplot are lower (Q_1) and upper (Q_3) quantile, median and the cut off point called fences, lie at the interval of $[(Q_1 - 1.5(IQR)), (Q_3 + 1.5(IQR))]$ where IQR stands for inter-quantile range obtain from the the difference between Q_3 and Q_1. Observations beyond the fences are considered as outliers. The extreme outlier happened when the data lie at the interval of $[(Q_1 - 3.0(IQR)), (Q_3 + 3.0(IQR))]$.

Histogram is another bar like graphical tool that is widely used in estimating the distribution of data. It can also be used to figure out the outlier. The data is said to be an outlier when a distribution is different from the bulk of data.

Kurtosis and skewness are methods which are used to characterise the location and variability of the data. Skewness is a measure of the distribution of the data. The value is considered zero when the distribution is normal. Positive value indicates that the data is skewed to the right and vice versa.

$$Skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}. \qquad (2.12)$$

where $\bar{Y}$ is mean, s is the standard deviation and N is number of data points.

Kurtosis is used to measure the peak or flat distribution. There are variety type of peak distribution which are platykurtic (<3), mesokurtic (=3) and leptokurtic (>3). Positive distribution indicates a peak distribution whereas negative distribution indicates a flat distribution.

$$Kurtosis = \frac{\sum_{i=1}^{N}(Y_i-\bar{Y})^4}{(N-1)s^4}. \tag{2.13}$$

where $\bar{Y}$ is mean, s is the standard deviation and N is number of data points. Skewness and kurtosis are also used for outlier detection. The rule of thumb says that a data is considered as outlier when skewness and kurtosis is fall outside the range of normal which is between -1 and 1 (Hildebrand, 1986).

Another simple method is simply converting the data point to z score and screen the absolute values (Donzenis & Rakow, 1987 studied by Jacobs, 2001). They suggested that z score of plus or minus 2.7 should be considered as outliers as the value is 1.5 times the interquartile range. In turn, if the z score of plus or minus 4.72, it should be considered as "far out" or in other word, it is called "contaminated outlier".

$$z_i = \frac{x_i-\bar{x}}{s} \tag{2.14}$$

where $s = \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right)^{\frac{1}{2}}$.

However, z score method is unsatisfactory especially for a small sample size because the $\bar{x}$ and s are not resistant since it is not unduly affected by a few unusual observations. Therefore Iglewicz and Hoaglin (1993) recommended modified z-score.

This method is more robust to the outliers as it relies on the median for calculating the z-score.

$$MAD = median_i\{|x_i - \tilde{x}|\} \tag{2.15}$$

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}. \tag{2.16}$$

Barnett and Lewis (1984) discussed about the "most extreme observation" in detecting the outliers. Extreme studentised deviate statistic test (ESD) is applied to detect the outliers in a random sample.

$$T_S = \max\{|x_i - \bar{x}| / s\}. \tag{2.17}$$

where $i = 1,2,\ldots,n$, s and $\bar{x}$ denote standard deviation and mean respectively. Assume x_j to the outliers. If the T_S exceeds the critical value, the x_j need to be taken out and process will be repeated for the rest of sample. Otherwise, the procedure is terminated. However this method might hide some extreme observations. This phenomenon is called "masking" in identifying the outliers. Details of discussion and illustrations are given by Iglewicz and Hoaglin (1993).

All the methods mentioned above focusing solely on univariate robust estimators and the extension version to multivariate problems is rarely discussed for several reasons. The huge size of data set, the complexity of the sample with many variables and the possibility of having missing value are among the obstacles where the univariate methods are capable to deal with.

Davies and Gather (1993) revealed that the outlier identification can be done through the specified lower bound, $L(X_N, \alpha_N)$ and upper bound, (X_N, α_N) where X_N is the random sample, $X_N = (X_1, X_2, \ldots, X_N)$ and α_N represent the number of outliers. All points either less than the lower bound or more than the upper bound are considered lying in the outlier region, $\text{out}(\alpha_N, \mu, \sigma^2)$ which can be written as follows:

Outlier region, OR

$$(X_N, \alpha_N) = (-\infty, L(X_N, \alpha_N)] \cup [R(X_N, \alpha_N), \infty). \tag{2.18}$$

The statistics lower bound, $L(X_N, \alpha_N)$ and upper bound, $R(X_N, \alpha_N)$ were proposed as below:

1. Mean and Standard Deviation

Let $\bar{X}_N$ denotes the mean and let $S_N = (\frac{\sum_{i=1}^{N}(X_i - \bar{X}_N)^2}{N-1})^{\frac{1}{2}}$ denotes the standard deviation of the sample, X_N. For some $g(N, \propto_N)$, we can identify all x satisfying to be $\propto_N$ outliers by the outlier identifier as

$$\left|x - \bar{X}_N\right| \geq S_N g(N, \alpha_N) \tag{2.19}$$

Thus, region of the outliers are

$$L(X_N, \alpha_N) = \bar{X}_N - S_N g(N, \alpha_N) \text{ and} \tag{2.20}$$

$$R(X_N, \alpha_N) = \bar{X}_N + S_N g(N, \alpha_N) \tag{2.21}$$

$$\text{where } \alpha_N = 1 - (1 - \alpha_N)^{1/N}. \tag{2.22}$$

2. Median(MED) and Median Absolute Deviation(MAD)

Hampel identifier yield as the follows

$MED(X_N) = (X_{\left[\frac{N+1}{2}\right]:N} + X_{\left[\frac{N}{2}\right]+1:N})/2$ and (2.23)

$MAD(X_N) = MED((|X_1 - MED(X_N)|, \dots, |X_N - MED(X_N)|).$ (2.24)

We can define an outlier identifier by having all x satisfying

$|x - MED(X_N)| \geq MAD(X_N)g(N, \alpha_N),$ (2.25)

Following these, region of the outliers are

$L(X_N, \alpha_N) = MED(X_N) - MAD(X_N)g(N, \alpha_N)$ and (2.26)

$R(X_N, \alpha_N) = MED(X_N) + MAD(X_N)g(N, \alpha_N).$ (2.27)

It has been shown by Hampel identifier that the latter provide a better identification of outliers. The distance measures between entities also used by many researchers to identify outliers. The famous Mahalanobis distance is

$MD_i = D_i(\bar{X}, S) = \{(x_i - \bar{X})^T S^{-1}(x_i - \bar{X})\}^{\frac{1}{2}}$ (2.28)

where $i = 1,2,3 \dots n$

is used by estimating the location and scatter a bulk of data where outliers are identified based on huge value (Hadi, 1992; Beguin & Hulliger, 2004). However, the problem of masking and swamping may arise. Small cluster of outliers can attract $\bar{X}$ and will inflate S in its direction and cause small value for MD_i. This is called as masking problem. Conversely, not all the observations with large MD_i value are necessary outliers. Small cluster of outliers can attract $\bar{X}$ and will inflate S away from some other observations which belong to the pattern suggested by the majority of observations. This is called as swamping problem. Penny (1996) comment on the critical value that use in MD_i and she proposed a better way when searching for a single outlier. Penny found that *Wilks's method* that recommended $\{p(n-1)/(n-p)\}/F_{p,n-p}$ is unsuitable and $p(n-1)^2 F_{p,n-p-1} / n(n-p-1+pF_{p,n-p-1})$ are correct critical value.

Koufakou at al (2008) proposed a new approach named MapReduce-AVF (*MR-AVF*) to detect the outliers for categorical dataset. *MR-AVF* is a parallel outlier detection method that is used to identify the outliers in a huge dataset. The user is not forced to devise a parallelization strategy for the task at hand but just require adapting it to a *MR-AVF* model. The map and reduce function are as below

$$map(k_1, v_1) \rightarrow (k_2, v_2)[] \tag{2.29}$$

$$reduce(k_2, v_2)[]) \rightarrow (k_2, v_3)[]. \tag{2.30}$$

First, the user defines key-value pairs, $k_1$and v_2 as input files. Then the user specifies what to do with the keys and values. A new output is produced with another set of k_2 and v_2. The reduced function sorts the key value pairs by k_2. Finally, all the associated values $v_2$are reduced and emitted as valuev_3.

Hadi and Simonoff (1993) created the first automatic method named forward search to deal with the multiple outliers in the data. The distance from the observed value y_i to fitted values can be calculated by

$$d_{i(m)} = \frac{| y_i - \mathbf{x'}_\mathrm{i}\, \hat{\beta}_{(m)} |}{\hat{\sigma}_{(m)} \sqrt{\{1 - \lambda_i \mathbf{x'}_\mathrm{i}\, (\mathbf{X'}_{(m)}\, \mathbf{X}_{(m)})^{-1} \mathbf{x}_\mathrm{i}\}}} \qquad (2.31)$$

where $\boldsymbol{X}_{(m)}$ denotes the matrix of $\boldsymbol{X}$, $\lambda_i = 1$ if the observation i is in the subset and $\lambda_i =$ -1 otherwise.

For the dimension, $p = 1$, Hadi and Simonoff (1993) suggested that the forward search should be stopped when the distance of $(m+1)th$ order is greater than $1 - \alpha/2(m+1)$ quantile of t distribution on $m - q$ degree of freedom.

In much larger dimension size of data where $p > 1$, Hadi (1994) used the square Mahalanobis distance

$$D^2{}_{i(m)} = (y_i - \hat{y}_{i(m)})' \hat{S}^{-1}{}_{(m)} (y_i - \hat{y}_{i(m)}). \qquad (2.32)$$

where $\hat{y}_{i(m)}$ denotes the fitted value for y_i which generated from estimate linear equation models. The estimate covariance matrix of the errors

$$\hat{S}_{(m)} = (m-q)^{-1} \Sigma \big(y_i - \hat{y}_{i(m)}\big)\big(y_i - \hat{y}_{i(m)}\big)'. \qquad (2.33)$$

Hadi (1994) suggested that this method should be stopped when it achieves $(1 - \alpha/n)$ quantile of the x^2 −distribution with degrees of freedom, p. The remaining $n - m$ observations are declared as outliers.

Grubbs (1950) categorized the causes of outliers as measurement errors, execution faults or intrinsic variability. Gross errors of measurement can yield outliers in a data set. No statistical method is required for that; such outliers can be weeded out without controversy. However, some outliers can be caused by unrecognized or execution error which cannot simply weeded out. Grubbs found that some initial model F should be specified in order to examine the outliers. If the outlier is discordant then model F must be abandoned as a homogenous model.

Chambers, Hentges and Zhao (2004) use the robust tree modeling to detect the presence of outliers for univariate and multivariate problems. WAID regression tree algorithm was used. There is no attempt to get the optimal trees. The splitting process of heterogeneity node was based on weighted sum of square residuals

$$\text{WSSR}_k = \sum_{i \in k} w_{ik} (y_i - \hat{y}_{wk})^2 \tag{2.34}$$

where $$\hat{y}_{wk} = \sum_{i \in k} \left(\frac{w_{ik} y_i}{w_{ik}} \right). \tag{2.35}$$

The weight w_{ik} is

$$w_{ik} = \frac{\psi(y_i - \hat{y}_{wk})}{y_i - \hat{y}_{wk}} \tag{2.36}$$

where $\psi(x)$ denotes the influence function.

For multivariate $\boldsymbol{y}$, Chambers, Hentges and Zhao (2004) proposed 3 options for building the regression tree for a p-dimensional response variable y which are average heterogeneity, average weight, and full multivariate. In the first option, WAID builts a tree by using the heterogeneity measure for a particular node, h at stage k where j and i represented the response variable and the case respectively.

$w_{ij}^{(hk)}$ denoted the weight

$$WRSS_{hk} = \sum_{i \in h} \sum_{j=1}^{p} w_{ij}^{(hk)} (y_{ij} - \hat{y}_{whj}^{(k)})^2 \tag{2.37}$$

where

$$\hat{y}_{whj}^{(k)} = \sum_{i \in h} \left(\frac{w_{ij}^{(hk)} y_{ij}}{w_{ij}^{(hk)}} \right). \tag{2.38}$$

By using WAID toolkit, these weight are based on robust influence function, the outliers' weight will be closely to 0 while the non-outliers' weight is approximately to 1. The proportion of error-generated outliers can be calculated by WAID. The optimal threshold value is

$$w^* = \arg \max_{w} [R_1(w)\{1 - R_2(w)\}] \tag{2.39}$$

where

$$R_1(w) \;=\; n_{error(w)}/N_{error} \text{ and} \tag{2.40}$$

$$R_2(w) \;=\; n_{non-errors(w)}/N_{outliers(w)}. \tag{2.41}$$

When two approaches (forward search and regression trees) are compared, the regression trees performed better than forward search.

Dynamic graphic seems has major potential. In future, this method will be ubiquitous. Becker, Cleveland and Wilk (1987) insisted that dynamic graphic methods have two important properties which are direct manipulation and instantaneous change of element. Haslett et al. (1991) discovered this concept for exploring and analyzing spatial data. This method can be used to examine local variability or so called anomalies in geochemical data. Identification of new tools in dynamic graphics is already recognized as for multivariate data analysis which seems to have major potential for spatially referenced data. In short, this method reflected the importance of map view. Map view has been overlaid on the schematic sketch of major components of stream network that has given rise to the data. The data of stream geochemistry (at region of Spain) contained 99 multivariate observations are available. The metals of *Pb, Zn, and Cu* are focused here. Variogram cloud and histogram have been used to identify regions of interesting variation in map view. Local anomalies are detected by using the scatter plot and variogram cloud.

Outlier detection has become an important topic in real world. Plenty type of outlier detection techniques have been created in order to trace the anomalies. Some practitioners prefer graphical methods or visual inspection; some practitioners opt to use data distribution. In fact, there is no the best techniques, it depends on the suitability of the case.

In this research, lower fence and upper fence in box plot are used to detect the errant observations by conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data. Every single point beyond the upper or lower fence is considered as outliers.

2.8.2 Outlier Handling

Dealing with the outliers is important part in an analysis. Discard or accommodate the outlier is a vital decision that should be made by the researchers. Simply ignoring the suspicious values could cause a huge influence in statistical analysis. Orr, Sackett and Dubois (1991) and Evans (1999) indicated that the outlier can be deleted straightforward when the individuals admit inattention during data collection or committed dishonesty responses. However, when the misappropriation cannot be justified, the techniques of handling must be applied for dealing with the outliers (Jacobs, 2001). The following section will discuss about several outlier handling methods.

Osborne (2002) suggested that outliers should not be removed. But, the outlier must be accommodated which could be done using transformation. This method is not only can reduced the skew and variance, it also keeps the extreme value in the data (Hamilton, 1992). However, transformation could alter the relationship between the original variables and the model. As a consequence, the scores might be hardly to interpret (Newton & Rudestam, 1999).

In particular, the researchers should consider the concept of robust method. Trimmed means and Winsorize means are among the popular estimators which are used to reduce the extreme value in the data (Barnett & Lewis, 1994; Jacobs, 2001). Both are less sensitive to outliers and give a reasonable estimate of central tendency.

1. Trimmed means

$$T = \frac{1}{n-2r}\sum_{i=r+1}^{n-r} X_{(i)} . \qquad (2.42)$$

2. Winsorize means

$$T_W = \frac{1}{n}\left\{\sum_{i=r+1}^{n-r} X_{(i)} + r\left[X_{(r+1)} + X_{(n-r)}\right]\right\} \qquad (2.43)$$

$$S_W^2 = \frac{\sum_{i=r+1}^{n-r}(X_{(i)} - T_W)^2 + r\left[(X_{(r+1)} - T_W)^2 + (X_{(n-r)} - T_W)^2\right]}{(n-2r)(n-2r-1)} . \qquad (2.44)$$

Confidence interval is

$T_W \pm t(1-\alpha/2)S_W$, in which $t(1-\alpha/2)$ comes from t distribution with $n - 2r - 1$.

In this research, we use the concept of Winsorize (Dixon, 1960) which the p% of data is simply removed from bottom and top of the elements and replaced by the remaining highest and lowest values. Wilcox (2005) recommended that 20% is the most suitable percentage in Winsorize process. However, the percentage, p can be determined by the researchers based on their own requirement or experience. This method has been chosen due to its less sensitivity towards the outliers but still provide a reasonable penalization on the data by replacing the given parts at the high and low end with the most extreme remaining values. At least, no outliers are excluded during the construction of tree model. Moreover, this method can reduce the magnitude of deviation and retaining its direction.

Sample:

$x_{i1}, x_{i2}, \dots x_{i100}$

Winsorize p% (let say $p = 5\%$)

Beginning Sample:

$x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6} \cdots x_{i95}, x_{i96}, x_{i97}, x_{i98}, x_{i99}, x_{i100}$

Winsorize Sample:

$x_{i6}, x_{i6}, x_{i6}, x_{i6}, x_{i6} \cdots, x_{i100}, x_{i100}, x_{i100}, x_{i100}, x_{i100}, x_{i100}$

2.9 Classification Tree

Classification and regression tree (CART) was introduced and popularised by Breimen et al. in year 1984 based on a recursive partitioning method suitable to categorical and continuous variables. This method has been widely improved and implemented due to its simplicity and transparency. The choice of this method in this study has been elaborated in Chapter 1.

Classification tree is a predictive modeling that has been widely used nowadays to predict the memberships of objects in the class of categorical dependent variable rather numerical value. The pseudo code is easy where

1. Start at a node (first node is called parent node).
2. For each X, find the set which minimize the sum of impurity in two nodes.
3. Find the split.

4. The tree is recursively partitioned into two child nodes until a stopping criterion is reached.

In order to get the best splitter, Gini index, Twoing and Entropy are generally used in CART for its impurity function in learning dataset.

i. Gini impurity index

$$Gini(t) = 1 - \sum_j [p(j|t)]^2 \tag{2.45}$$

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(i) \tag{2.46}$$

where $p(j|t)$ is the relative frequency of class j at node t, k is the number of children nodes, n_i is the number of records at child i and n is the number of record at node p.

ii. Twoing

$$\frac{p_L p_R}{4} \left(\sum_i (|p(i|t_L) - p(i|t_R)|)\right)^2 \tag{2.47}$$

where L and R are left and right sides and $p(i|t)$ is relative frequency of class i at node t and p_L and p_R represent the probability in left node and right node respectively (Breimen, 1996).

iii. Entropy

$$-\sum_i p_i log_2 p_i \tag{2.48}$$

where p_i is the relative frequency of class i at node t.

Twoing is resembles Gini index if the target group is binary. However, for multi-class problem, it prefers attributes with evenly divided splits. In comparison between Gini and entropy, Gini is intended for continuous and categorical attribute whereas entropy intended for categorical only but some modification have been done so that it can also be used for continuous attribute. Gini tried to get the largest class instead of finding groups that make up 50% of data as in entropy. It does not involve any logarithm computation as in entropy. Also, Gini tends to gain the minimize error but entropy is an exploratory analysis where it summarises main characteristic (often visual methods) which can tell us beyond the formal modeling or hypothesis testing. Gini is opting to be performed in CART whereas entropy (info gain) is more favorable in C4.5 and ID3 (Zambon, Lawrence, Bunn & Powell, 2006; Apte & Weiss, 1997; Raileanu & Stoffel, 2004).

In fact, it is not obvious which of them produce the best decision tree. Large amount of empirical tests were conducted by Raileanu and Stoffel (2004) to determine which measurement produces better result. However, there is no conclusive result as only about 2% differences between them. But due to the suitability with the proposed algorithm, Gini index has been used in this research. Some modification on Gini index measurement is implemented in order to find the best splitting point when confront the existence of outliers in the data.

One of the advantages of tree is it can isolate the outliers without the need of taking of the outliers in the data. However, many researches have been proved that tree should go through a pre-processing stage. In other word, outlier must be handled well before constructing it so that the accuracy of tree is not affected. Also, handling outlier can avoid tree to become too bushy which might produce an unrealistic tree. Besides, based on the examples given in Chapter 1, it can be seen that the outliers affect the sensitivity of tree. It means that some of the useful variables might be masked by the useless variables due to the influence of outliers. In fact, outliers affect the Gini index measurement; the cutting point could be shifted due to the heavy tail in the data.

To evaluate the performance of tree, cross validation or hold out validation are commonly used to estimate the error rate. Of course, looking at the error rate itself is insufficient to measure the performance of a constructed tree hence the structure of tree, number of leaves, number of splits and other criteria must be also taken into account. A tree with a very low error rate but having a bushy tree is considered unsatisfactory.

2.10 Pruning Methods

Overfitting is a common issue in tree. It happens when the learning algorithm continues to develop the branches of tree to its maximum. If the tree is fully grown then it loses out its generalization capability. There are few causes of overfitting tree. For instance, it is caused by the presence of noise in data, lack of representative instances, failure to compensate for algorithms that explore a large number of alternatives and so forth. Therefore, pre-pruning or post-pruning approaches should be used to avoid overfitting tree.

Pre-pruning means the tree will stop growing before it is fully grown. Generally, the method is more likely to be implemented in CHAID which is done from top to bottom. The stopping criteria are vital to stop the condition for a node. In more restrictive conditions, the tree stops growing when it reaches user-specified threshold or it stops when the class distributions of instances are independent of the available features.

Meanwhile, post-pruning means the tree is growing to its maximum. Then, the pruning process (Breimen et al., 1984) was developed to cut back the branch of tree either bottom-top or top-down transversal of the nodes which the removed branches are not contributing to the generalization accuracy. There are many studies have been carried out and proved that pruning process can reduce the effect of noisy domain and increase the accuracy (Bratko & Bohanec, 1994). The followings describe some popular pruning techniques.

i. Cost-Complexity Pruning

This technique is commonly used in CART where the error error of a tree based on the test set plus a penalty factor for the size of the tree (Breimen et al., 1984; Rokach & Maimon, 2008). The more leaves contain in the tree means that the higher complexity in the tree due to more partitioning of the data into smaller pieces and more possibilities for fitting the training set. Generally, the basic idea of cost complexity pruning is only consider those with the "best of their kind" in the sense below instead of consider all pruned sub trees. Total cost-complexity measure, $R_\alpha(T)$ of tree T is defined as $R(T) + \alpha|\tilde{T}|$, where $R(T)$ is the fraction of validation that misclassified by tree, α is complexity parameter which is adjustable and $|\tilde{T}|$ is the number of leaf in node T.

Initially the maximum tree has no error, but replacing the sub trees with leaves increase the errors. The idea of this method is to calculate the number of errors for each node if collapsed to leaf compare to the leaves which taking into account more nodes used. The α is calculated for each node and the node with smallest α branch will be pruned. Repeating the process for the sequence of trees $T_0, T_1, T_2, \dots, T_k$ then pick the one with the minimum error error on the test set.

Most of the traditional tree applied this method in pruning method as it is the state of art in CART which introduced by Breimen et al (1984).

ii. Reduce Error Pruning

This method was proposed by Quinlan in year 1987. During the pruning process from bottom to top, the procedure is to check the accuracy of tree when it is replaced by a most frequent class. If reduced tree does not reduce the accuracy the node can be pruned. The process is repeating till the pruning process decrease the accuracy. At the end, the tree produces a smallest version with accurate subtree.

iii. Rule Post-pruning

This approach is commonly used in C4.5. This method converts the tree into rule (one for each path) and then examines the rules with the purpose of simplying them without losing any accuracy. The rule will be removed if the error rate on the test set does not decrease. Finally, sort the final rule into desired sequence for use (Bramer, 2013).

iv. Critical Value Pruning

Critical value pruning was proposed by Minger (1987) where it is also a bottom-up pruning technique which tries to collect the information gain during the growing of tree. Recall that the splitting criteria during the growing of tree, information gain has been used for the measurement so that the purer subset can be obtained. The measurement reflects how good the selected attribute split the data between the groups in the data. This technique specifies a critical value and prune those nodes that do not reach the critical value unless further along the branch reach it. When the subtree is considered to be pruned, the value of splitting criteria needs to compare to the threshold and if the value is small then replace the tree by a leaf.

In fact, there are still many pruning approaches such as minimum descriptive length (MDL) pruning, optimal pruning, minimum error pruning (MEP), pessimistic pruning, error based pruning (EBP) and so forth (Rokah & Maimon, 2008; Frank, 2000).

2.11 Pre-processing and Its Drawback

Current practice on the process of outliers' detection and missing value imputation are normally gone through separately with the process of constructing the classifier. This means that initially the contaminated data will be sorted, filtered, and solved in order to get the "pure" (without outliers) data. Then, the "pure" data will be used to construct the classifier. The test set will be used to evaluate how accurate the model is by considering the error rate. Finally the model is used for prediction. However, the current method is not protecting outliers especially when it contains outliers due to its rejection from the early stage. Once the data is removed, it will no longer used in the data. According to Engel and Theusinger (1998), having a clean data clearly is too academic and not realistic especially in real world application (Engel, Evans, Hermann & Verdenius, 1997). As we know that outliers can be legitimate or illegitimate. If it is illegitimate, removing the outliers can produce desirable outcome. In contrast, if legitimate outliers are removed then it is considered bias as it is unlikely to be representative to the whole population (Orr, Sackett, & DuBois, 1991).

In fact, outlier is considered too important in certain field such as in computer networking, medical field, banking and so forth. The application of current mechanism is not suitable to them as no protection is given to the data once the classifier is developed. Let us look at some examples here.

In medical field, despite health profession are well developed over the last few decades, the case of medical error is still a serious issue that keep happening. 44,000 to 98,000 of the Americans die every year due to the medical errors based on a report to *Err Is Human -- Building a Safer Health System* (Kohn, Corringan & Donaldson., 2000). And, the cost of injury due to the medical errors is about 17 billion dollars (Thomas et al., 1999). In this case, outliers or anomalies are vital as they can be used to monitor the data driven and alert the framework based on the patient clinical record. Hauskrecht et al. (2010) developed a data driven approach from electronic health record (EHR) to detect the unusual patient management decision which can lead to true alert rates without constructing alerting model from the experts.

Besides, network intrusion is another issue that received a lot of attention from all fields especially computer network security. The malicious activity is performed, trying to hack and spread viruses, Trojans and worms into the local and remote machine. To detect various type of attacks, outliers is the vital information to protect the network whilst to reduce the false alarm rate. Therefore, outlier is too powerful in solving the real world problem. Removing them during the pre-processing stage will produce a pure classifier but is it reliable to be used for future classification or prediction? So, creating a natural technique which resist to outlier itself from level to level is extremely important in order to create an accurate classifier for prediction.

In the past two decades, many works have been done by researcher to refine the issue of handling outlier in tree. John (1995) showed a new approach which was to rebuild the tree using the reduced training set. The reduced training set means that the original training set minus the pruned sub tree. The pruned sub tree was considered as un-informative records or outliers. Then, retrain it to construct a new tree. This method is good as all records were included and fewer nodes can be produced with high accuracy. However, it might create an extremely bushy tree and might be wasting time to examine the difference in number of nodes between a tree built with and without the set of points. Moreover, some of the outliers could be meaningful too. Removing the outliers could bring an inaccurate model for prediction. To generate a good decision tree, pre-processing is vital to improve the data description. Terabe, Katai, Sawaragi, Washio and Motoda (1999) commented that most of the pre-processing takes mush running times. And all based on logic programming with a need of priori knowledge. They proposed an association rules which can perform even better and priori knowledge can be neglected. By having the antecedent (if) and consequent (then), new attribute can be generated. Then, it is used to construct a tree. According to the result, decision with pre-processing showed more effectiveness than the decision tree with original data. Rajendren, Madheswaran and Naganandhini (2010) applied this idea on brain tumor diagnosis from CT scan brain image for tissues abnormalities detection, sharp analysis and so forth. The process of CT scan brain image was called shape priori technique. The general idea of shape priori technique is to evolve a curve of the image for the shape segmentation which has given the efficient features to be stored in transactional database. Then, association rule is used to mine the features which were acceptable for classification task. Finally decision tree was used for classification and categorization. The proposed method (association rule mining (ARM) + decision tree) showed a better performance compared to traditional association rules and neural network. As we know, it is not easy to discover interesting relations between variables especially dealing with a small data set. Besides, it takes time to merge the attribute. Some information could be also masked when new attributed is generated. As in shape priori technique, the prior knowledge of the shape of curve is required which might be the constraint to the practitioners.

Nowadays, dealing with a large collection of spatial data is inevitable and it is crucial to index them to support the process of query. R*-tree has implemented into commercial system and performed quite well. However, improving R*-tree is still needed in outliers identification and storage at higher levels of the spatial tree index. R^0-tree is the one to be used to improve the performance of R*-tree where outliers were stored at higher levels with smaller minimum bounding rectangles at lower-level nodes which performs much more better. There are 5 spatial query were implemented and the results showed that R^0-tree were significantly outperforms in all cases (Xia & Zhang, 2005).

Muniyandi, Rajeswari and Rajaram (2011) used k-means clustering to partition the training instances into k-cluster using Euclidean distance similarity. This method is implemented to solve the intrusion problem in network environment. The separation of normal and anomaly region are used to build C4.5 and this method performs the best among classifier. This method is performed well as it leads to the highest precision and accuracy rate. Decision tree based on the idea of clustering that resembles this method has also been used in HMM-based speech synthesis techniques with the criterion of maximum likelihood (ML) or minimum description length (MDL). However, due to the sensitive of ML or MDL towards the outliers (discrepancies), the trees performed poorly where optimal clusters are not achievable.

Kyung, June, Dao and Nam (2011) proposed an algorithm which outliers must be detected and removed. By comparing between 'conventional', 'no preference' and proposed' method, the proposed method performed the best which mean in a sentence, the proposed decision tree based clustering with outlier removal produced a well-balanced speech quality. Using clustering is good but the practitioner should have to know what the clustering is all about. Moreover, outlier might be sometime meaningful; simply removing is not a good way as it brings to bias classifier. Time consuming is one of the problems too in this method.

In other away round, decision trees and data pre-processing also been used by Parisot, Ghoniem and Otjacques (2014) to help clustering interpretation. This study proposed an evolutionary algorithm to pre-process the data using transformation of data so that the transformed data can be more easily to be interpreted and yield a simpler tree. The clustering of transformed data set lead to a smaller size in tree with lower error rate. Even though this method showed good enough in the end result but some potential variables might be masked during clustering. Even some features have been mixed to form the cluster which might uninformative during the construction of tree.

Local outlier factor (LOF) has been used to measure the local diversity by using distance to calculate the density. Fawagrh, Gaber and Elyann (2015) has proposed an out performed method named LOFB-DRF to improve the random forests in pruning level. This method selects the diverse trees in RF then used the trees to form a pruned ensemble of the original one. And, it showed that LOFB-DRF perform high accuracy to 99%. This idea is superb as it balanced up the size of tree too. This method is great but again, the knowledge of clustering is required during the selection of highest weighted LOF value. This could bring some troubles to the practitioner who is not from classification background.

To avoid noise data, Wang, Gu and Wang (2014) have introduced another way to get a more robust ID3 tree by using insensible attributes as priority instead of sensible attributes. The results from few data showed that the accuracy of insensible method is higher compared to sensible method. Therefore, decision tree induced by insensible tree is more robust than others. However, by selecting the "most unimportance" attribute as priority could create a big size of tree. Moreover, tree has its transparent nature, it explicit all the possible alternatives so that we can easily traced back the entire useful attribute along the process. However, using the insensible attribute could increase the ambiguity in decision making.

After discussing some previous research, we found that some methods can really perform well but some are not realistic in real world application. Most of the methods are not really accomodating the outliers but focusing more on the end result of the tree. High accuracy of tree from a pure data is seemed like too common in many studies. The data will be scanned in the pre-processing stage. It means that all the contaminated data will be detected and penalized before entering to the next stage. Obviously, we can definitely get a "clean" or "pure" data which will be used for constructing the tree. And, the tree created by the "clean data" will surely provide an "overconfident" estimation that might lead to high significant error. Consequently, the constructed model for prediction will definitely bias no matter how good the final result is. Outliers must be investigated further since it can be due to data entry error, incorrect distribution assumption, in fact valid or other factors. It is meaningless if the tree constructed by excluding all the inevitable outliers.

Substantially, in hospital, outlier is vital for diseases diagnosis or any pattern recognition. However, doctor has limited time to go through the huge historical profile of all patients. Some studies seem unreliable as it takes a long time for a doctor to go for sorting, inspecting, analysing and interpreting. In banking, the variety task of customer profiles is increasing rapidly. The powerful automated decision support system is needed not only for prediction but it must be able to identify the fraudulent based on the anomaly in the data. Besides, the information of outlier is also vital to help the sector to decide whether to approve the credit card application since the number of bankruptcy is increasing recently. Some of the company would even hire an expertise to handle millions of data by sorting out all the anomalies. But this method demands expensive processing time and the cost of handling the bad data can be enormous. Even, when all the data have been filtered, the ultimate data will provide us a very pure dataset which might provide us an unsubstantial estimation. Thus, they need an extremely reliable way to handle the outliers, model construction and prediction simultaneously without removing any of the outliers in the data. This study proposes a new mechanism on constructing a tree that penalising the occurrence of outliers during Gini purity measurement. It offers better way for practitioners for using the tree without burden too much on the occurrence of outliers. Details of the study will be elaborated extensively in Chapter 3.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

Classification and regression tree (CART) has been proven works satisfactorily in some classification problems where some given studies have been discussed in Chapter 2. However, the successful process of classifying objects using tree is influenced by the structure of data. CART is very powerful and suit for data mining tasks as straight forward relationship between the variables that goes unnoticed can be revealed instead of using much complex methods. A series of if-then statement manages to classify observations in a particular manner especially in business problem. However, one of the challenges in dealing with continuous variables is the possibility of the occurrence of outliers. The used of data with outliers may lead to the constructed of bias CART hence end with misleading results.

This study initiates on constructing a CART that is able to accommodate wisely the occurrence of outliers along the construction process in order to achieve an unbiased classification process. The proposed CART is believed to give a better offer to practitioners in analysing huge data and small data sets when the process of cleaning is impractical to be done due to some constraints such as limited time for analysis, shortage manpower knowledge, expertise and etc. This chapter discusses extensively the idea on constructing a CART that insensitive towards the occurrence of outliers.

3.2 Framework of Study

The proposed CART offers an alternative method for practitioners in classification problems when the data is believed contaminated with true outliers. The idea of this proposed CART lays on the strategy that simultaneously handles and accommodates the outliers during the process of constructing the tree. It gives an advantage to practitioners to directly use the method rather than taking a preamble analysis to clean the data prior to constructing the tree. In general, the framework of the classification as proposed in this study are mainly separated into 5 parts which are

i. Data inspection – The data is investigated for the existence of outlier.

ii. Outlier handling – The detected outlier will be handled by using Winsorize method.

iii. Gini purity measurement and tree construction – Each variable is computed to identify for goodness of split. The selected variable will be used as the splitting attribute.

iv. Stopping rules – Three stopping rules have been introduced in this study to stop the tree from being too bushy.

v. Evaluation - Error rate and tree size are used to measure the performance of tree.

3.2.1 Data Inspection

Most methods for identifying outliers as discussed in Chapter 2 are focusing on penalising the values so that it leads to the lowest variance and other statistics measures. Such aims may not be practical in classification problems when the whole focus is to ensure that the constructed rule is capable to allocate a future object to its correct group, i.e. minimise the error rate and not directly focus to the statistics. Thus, a careful selection of method for outlier penalisation must be consistent with the aim for constructing a CART. This study has considered lower and upper fences in a constructed boxplot as the indicator to identify outliers in the data. This strategy would enable us to determine which objects with particular variables are potential outliers. A boxplot is constructed for each variable *j,* for *j=1,2,...,p,* by sorting the values in ascending order. Then, we determined the median of variable *j* at the 50th percentile of the data, the point at 25th percentile, the point at 75th percentile and the range between the 25th and 75th percentiles of the variable. Thus, the lower fence of variable *j* is obtained by the following equation

Lower fence: $L_j = Q_{1j} - a \times IQR_j$ (3.1)

and the upper fence of variable *j* is

Upper fence: $U_j = Q_{3j} + a \times IQR_j$ (3.2)

where Q_{1j} is the first quartile (or 25th percentiles) of variable *j*, Q_{3j} is the third quartile (or 75th percentiles) of variable *j*, IQR_j is the different between Q_{3j} and Q_{1j} and a is a constant that set the wide of these fences.

The fence points as given in (3.1) and (3.2) give us information about limitation of points under the "normal" of data and indirectly highlight possible outliers in each variable *j.* We mark a value of variable *j* as outlier if it is less than the lower fence, L_j, or if its value is greater than the upper fence, U_j.

3.2.2 Outlier Handling

Generally, it is easy to handle the outlier by simply removing the identified value. However, such action is arguable as it may alleviate the actual behaviour of the variable. Eliminating outliers can only be considered if there is evidence that the value is recorded wrongly. Therefore, this study has considered a wise method called Winsorize method which the outlier is penalised and retained in the data without removing them.

Winsorize is a method that replaces the lowest and highest values (outliers) with observations closest to them. Instead of eliminating the outlier, the observation is altered, allowing for a degree of influence. Let n be the number of observation of a training set. Generally, the training set and the test set are set to be $\frac{2}{3}n$ and $\frac{1}{3}n$ respectively. Let $X_k = \{x_{1k}, x_{2k}, \dots, n_{nk}\}$ be a variable that has been identified having outliers using the boxplot as described in section 3.3.1 and 10% (based on own suitability) be the percentage of penalisation to the training set of variable X_k. Then, the number of object on the both side of tail that needs to be Winsorized is determined via

$$\alpha = 10\% \times \frac{2}{3}n \tag{3.3}$$

$$X_k = \{x_1, x_2, x_3, x_4, x_5, \dots, x_{n-3}, x_{n-2}, x_{n-1}, x_n\} \tag{3.4}$$

If for example, $\alpha = 3$, then we altered the data such that the Winsorize data $X_{wk} = \{x_4, x_4, x_4, x_4, x_5, \dots, x_{n-4}, x_{n-3}, x_{n-3}, x_{n-3}, x_{n-3}\}$ is obtained. The Winsorize data is used to compute a splitting point of the variable.

Figure 3.1 showed an example of data in the process of detecting and winsorizing. Based on the computation, for PA500, 0.05 is considered as lower fence. Therefore, all the values below 0.05 are considered as outliers which Winsorize method needs to be carried out.

	PA500 (Original)	**PA 500 (Winsorize)**	**Group**
Outliers	0.01	0.05	adi
	0.02	0.05	fad
	0.03	0.05	con
	0.04	0.05	con
	0.04	0.05	con
	0.04	0.05	fad
	0.04	0.05	adi
	0.04	0.05	fad
Lower fence	0.05	0.05	fad
	0.05	0.05	mas
	0.05	0.05	con
	0.05	0.05	adi
Normal data	0.06	0.06	mas
	0.06	0.06	con
	0.06	0.06	gla
	.	.	.
	.	.	.
	.	.	.

Figure 3.1. Arrangement of data before and after winsorizing

3.2.3 Gini Purity Measurement and Tree Construction

CART involves the process of partitioning the data sets into levels. Every single split will be based on the splitting criteria. In order to determine the best variable for splitting the data, some measurements are needed that would allow us to compare the variables on some scales and choose the highest among the other. We used Gini purity index as our measurement which means that we are focusing on the highest purity level (lowest impurity).

Once the data has been inspected and handled, the data in each variable will be sorted for Winsorize Gini purity index is computation. The splitting point is chosen based on the class of paired that hold greater number of objects. The splitting point (SP) which provides a maximum homogeneity (highest Gini purity) for the node will be selected as the splitting variable and splitting point.

Following Gini purity index by Breiman et al. (1984), we used the function on the Winsorize data. Therefore, we obtained Winsorize Gini purity index as

$$G_w = \sum_j [p(j|t)]^2 \tag{3.5}$$

where $p(j|t)$ is the relative frequency of class j at node t. Then, the weighted average of Winsorize Gini purity index is

$$WA_w = \sum_{i=1}^{k} \frac{n_i}{n} G_w. \tag{3.6}$$

The highest Gini purity between the variables will be selected for that particular node. Figure 3.2 shows an example of Gini purity measurement after winsorizing.

PA 500 (Winsorize)		Group											
		≤ 0.05						> 0.05					
		adi	car	con	fad	gla	mas	adi	car	con	fad	gla	mas
		3	0	4	4	0	1	13	16	6	5	14	14
0.05													
0.05													
0.05													
0.05													
0.05													
0.05	0.05												
0.05													
0.05													
0.05													
0.05													
0.05													
0.05													
0.06													
0.06													
0.06													
.													
.													
.													

$$G_{w_{left}} = \sum_j [p(j|t)]^2 \qquad G_{w_{right}} = \sum_j [p(j|t)]^2$$

$$WA_w = \sum_{i=1}^{k} \frac{n_l}{n} G_{w_{left}} + \sum_{i=1}^{k} \frac{n_l}{n} G_{w_{right}}$$

Figure 3.2. Winsorize Gini purity computation

Goodness of split criteria is the decrease in impurity:

The maximum purity measure is

$$\Delta i(t) = 1 - [-\sum_{k=1}^{k} p^2(k \mid t_p) + p_l \sum_{k=1}^{k} p^2(k \mid t_l) + p_r \sum_{k=1}^{k} p^2(k \mid t_r)] \tag{3.7}$$

$$\Delta i_w(\delta, t) = 1 - [i_w(t) - P_L i(t_{wL}) - P_R i(t_{wR})] \tag{3.8}$$

$$\underset{x_j \le x_j^R, j=1,\ldots,M}{\arg\max} \; (1 - [-\sum_{k=1}^{k} p^2(k \mid t_p) + p_l \sum_{k=1}^{k} p^2(k \mid t_l) + p_r \sum_{k=1}^{k} p^2(k \mid t_r)]) \tag{3.9}$$

$$\Delta i_w(\delta^*, t) = \underset{\delta \in S}{max}\, \Delta i_w(\delta, t). \tag{3.10}$$

where $\Delta i_w(\delta^*, t)$ is the goodness of split and split is denoted as δ. Such an ongoing process will solve the maximization problem at each node.

Let t_p be the parent node and will be separated into left, t_l and right, t_r nodes respectively based on the selected variable, x_{kw} and splitting point. The maximum homogeneity of left and right nodes will be equivalent to maximum decrease of impurity as shown in Figure 3.3.

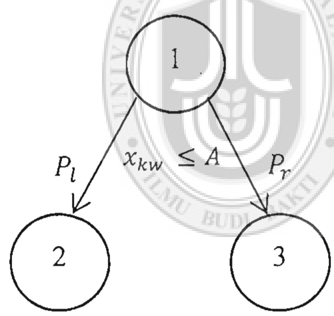


Figure 3.3. Goodness of split

3.2.4 Stopping Rules

The growing of tree continues until a stopping rule is triggered. This study uses three rules to avoid from creating a bushy tree. A post-pruning method which is used to handle the loosely stopping criteria can also be avoided. The node stops splitting when it reaches one of the following thresholds:

1. The node contains 70% or above of homogeneity.
2. The node meets the minimum observation, n_{min} , which being set as to have 10% or 15% of total observations, *N*.
3. If the computed Winsorize *Gini* purity index within and between variables are equal or greater than 70%, the node will have its final split called terminal nodes. In fact, the proportional of the thresholds can be adjusted based on the practitioner's needs. The higher the proportional set, the higher the accuracy of the tree classifier. However, bushy tree could be produced in the end. In contra, setting too low on the proportional could be a risk to classify the future objects.

Threshold 2 is set following Rokach and Maimon(2008), while thresholds 1 and 3 innovated the idea of Breimen (1984) and Kantardzic (2011). Small study was conducted and presented in Section 4.2 in an attempt to identify the stopping percentage.

All the child nodes or non-terminal nodes need to be inspected using the threshold 1 and threshold 2. If the non-terminal node contains 70% of homogeneity or meets the minimum number of observations, then we can stop the process of splitting and assume the node as terminal node (final node).

During the splitting process, Gini purity index is computed in order to choose the best splitting variable and splitting point. If the variable gains 70% or above of the Gini purity index measurement within and between the variables as in threshold 3, then we can conclude that the splitting point has been successfully split of group up to its maximum homogeneity which considered as its final split. More details are explained in Chapter 4.

3.2.5 Evaluation

The true error rate, $R^*(d)$ is used to estimate the accuracy of a classifier. In this study, test sample estimation is used which the observations from the learning set, L are divided into two sets L_1 and L_2. The observations in L_1 are used to construct the model, d. The observations in L_2 are used to estimate the error rate, $R^*(d)$. If n_2 is the number of observations in L_2, then the test sample estimate, $R^{ts}(d)$ is defined by

$$R^{ts}(d) = \frac{1}{n_2}\sum_{(x_{n,}\, j_n)\in L_2} x(d(x_n) \neq j_n). \tag{3.11}$$

where L_1 is training set and L_2 is test set.

Then the error rate is merely the proportion objects being misclassified by the constructed tree. Lower error rate indicates good performance of the tree.

3.3 Tree Algorithm

The outlines for the whole processes as discussed in Sub-sections 3.2.1 to 3.2.5 is summarised in an Algorithm 3.1:

Algorithm 3.1

Winsorize Tree Algorithm

Step 1	Get the data ready. Split the data into two mutually sets called training set and test set. Let 70% of the data in the training set and 30% of the data in the test set for inspection.
Step 2	Based on the training set, construct a Boxplot. Then, use upper fence and lower fence of the Boxplot to check on the present of outliers for all the variables respectively.
Step 3	Arrange the identified continuous variables with outliers from step 2, in order.
Step 4	Winsorize each variable as follow: Step 4.1: Determine the splitting point by measuring the Gini purity index. Step 4.2: Compute the Gini purity index using Winsorize Gini purity index at each splitting point. Step 4.3: Choose a splitting point on the class of paired that hold greater number of objects.
Step 5	Compare the highest Winsorize Gini score between the variables: Step 5.1: Choose the variable that scores the highest Winsorize Gini. This is called the goodness of split which provides the highest homogeneity. Step 5.2: Check for stopping rules. 5.2.1: If the computed Winsorize Gini purity index within and

between variables are equal or greater than 70%, the node will have its final split called terminal node.

5.2.2: Else if, the split nodes is still considered as non terminal node unless the node reaches 70% or above of homogeneity or it reaches its minimum observation, n_{min} .

5.2.3: Else, repeat from Step 2.

Step 6: Use the test set to compute for the error rate.

Step 7: Print the error rate.

For easy view, Algorithm 3.1 is presented in a flowchart form as in Figure 3.4.

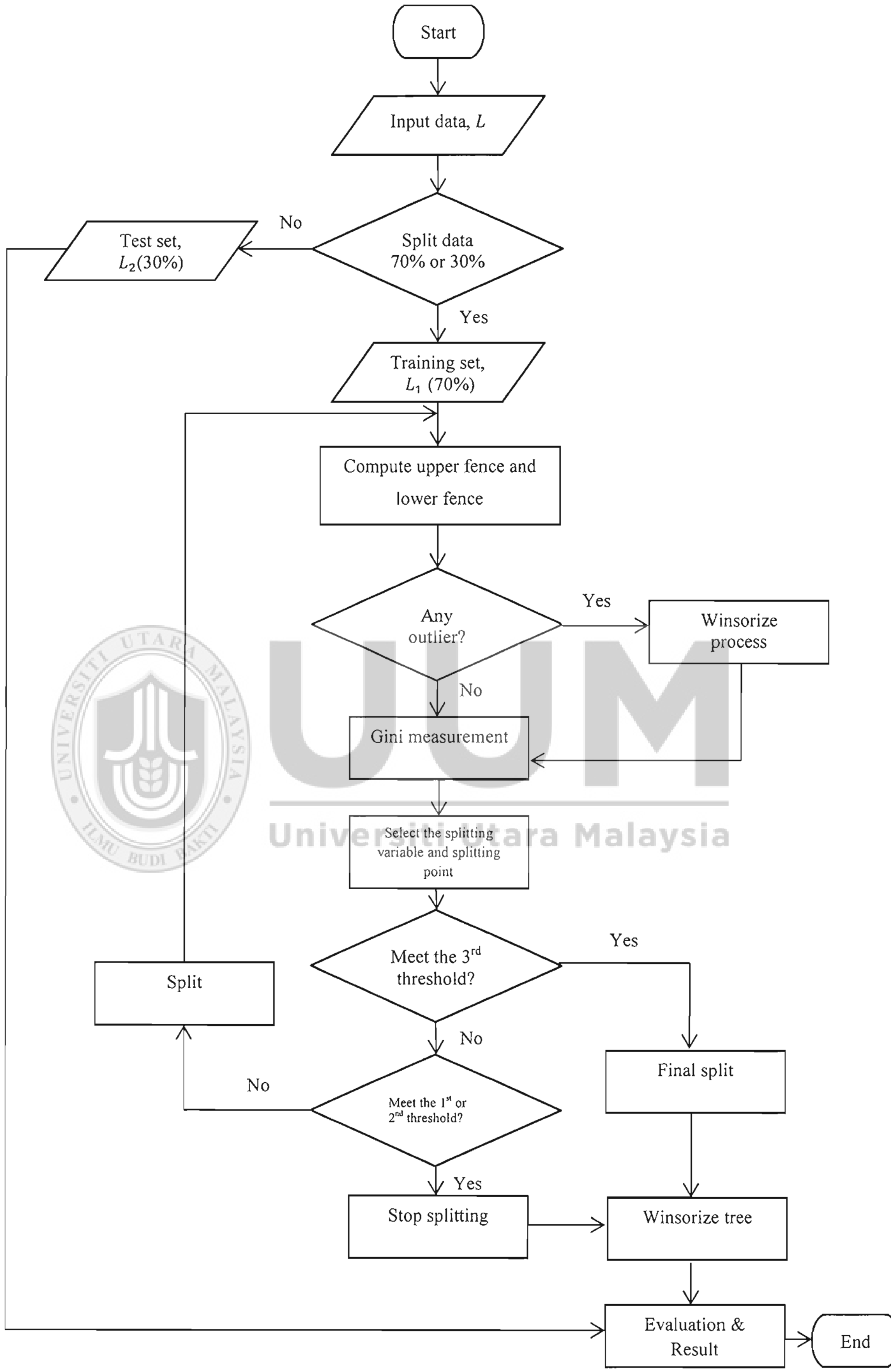


Figure 3.4. Flow chart of Winsorize algorithm

3.4 Data

Seven real data sets have been chosen for this research. The selected data are categorised as small, medium and big data sets from different background. The purpose of applying different size of data is to get evidence how the proposed is comparable to the traditional tree. For example, people used tree in prognosis scoring for cancer outcome predictions. Besides, it is also allowing decision-makers to apply evidence-based medicine to make objective clinical decisions when faced with complex situations. To prove that our propose method is comparative to the traditional tree in all areas; we also try some data from other fields such as life and archaeology. Data that we used are named *Breast Tissues* (Jossinet, 1996; Silva, Marques & Jossinet,2000), *Egyptian Skull* (Egyptian Skull Development. (n.d.). *StatLib and Story Library.* Retrieved June 2014, from http://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html), *Pima Indians* (Smith, Everhart, Dickson, Knowler & Johannes, 1988), *Iris* (Fisher, 1936; Duda & Hart, 1973), *Bumpus Sparrow* (Bumpus, 1898), *Indians Liver Patients Data (ILPD)* (Jayakrisharan, Rajan, Jagdish & Sanjay, 2014) and *Kyphosis* (Chamber & Hastie, 1992). *Bumpus Sparrow* and *Kyphosis* are considered as small data sets while *Egyptian Skull*, *Iris* and *Breast Tissues* are considered as medium data sets. And, *Pima Indians* and *ILPD* are considered as big data sets. More descriptions of the data are given in Chapter 4 where the details and results are explained according to each case.

Table 3.1.

Data Description

Data	Size	Number of group	Number of variables	Total number of observations
1. Bumpus	Small	2	6	49
2. Kyphosis		2	4	81
3. Breast Tissue	Medium	6	9	106
4. Egyptians Skull		2	5	150
5. Iris		3	5	150
6. ILPD	Big	2	11	583
7. Pima Indians		2	9	768

All the computation for completing the whole process are performed using Windows 7 Home Premium with processor of Intel (R) Core (TM) i5-2450M CPU @2.5GHz and 4GB of RAM. R software version R 2.12.0 has been used to run the whole analysis.

CHAPTER 4
ANALYSIS

4.1 Introduction

This chapter discusses on the analyses of the proposed Winsorize tree carried on some real data sets. As have been outlined in Chapter 3, the proposed tree starts by screening a data set using the box plot in order to identify any possibility of outliers. Then, the variable with the identified outliers is Winsorized so that the computation of Winsorized Gini purity index would not be affected by the outliers. We chose the variable with the highest Winsorized Gini purity index to be split, which led to new branches. These processes are repeated until one of the three stopping rules is met, as discussed in sub-section 3.2.4. We investigated the performance of the proposed classification (Winsorize tree), on seven well known data sets namely *Breast Tissue* (Jossinet, 1996; Silva, Marques & Jossinet, 2000), *Egyptian* (Hand, Daly, Lunn, McConway & Ostrowski, 1994), *Sparrow Bumpus* (Bumpus, 1898), *Pima Indians* (Smith, Everhart, Dickson, Knowler & Johannes, 1988), *Iris* (Fisher, 1936) and *Indians Liver Patient Dataset (ILPD)* (Ramana, Babu & Venkateswarlu, 2012) and *Kyphosis* (Chamber & Hastie, 1992). All the data stated above can be retrieved from UCI machine learning repository. Each data was investigated following three stages: (i) we conducted preamble analyses based on descriptive statistics and univariate groups comparison test in order to get an early information about the behaviour of the data, i.e. existence of outliers, distribution of the data and behaviour of variables which include an ability of variables to discriminate the groups, (ii) we constructed the proposed Winsorize tree using a training set and finally (iii) we used a test set to evaluate the constructed tree in order to measure its performance.

Also, we performed traditional tree and pruned tree to allow for performance comparison purposes based on the computed error rate. The traditional tree and pruned tree is following the idea of Breimen (1984) which details on these trees have been outlined in Chapter 2, Section 2.9 and Section 2.10 respectively. Besides of giving full concentration of the performance of the trees based on the error rate, our discussion also focuses on each component used in constructing a tree. The discussion touches on the effectiveness of the box plot used for identifying the outliers in multivariate case, the usefulness of the Winsorize approach in estimating the purity of the data in each node (Gini purity) for splitting process and the workable of the proposed stopping criteria for stopping the tree recursive process from being bushy.

4.2 Identifying Percentage of Homogeneity for Stopping Rules

In fact, choosing a significant percentage in stopping rules is vital so that the tree is neither under fitting nor over fitting. As discussed in Section 3.2.4, there are three stopping criteria to stop the tree from further splitting. The splitting stops when the relative node reaches the relative decrease in impurity (increase in purity). Suppose the tree will stop when there is a single observation in each child node or all the observations within each node are identical distribution of predictor variable. However, these thresholds seem hardly to be achieved in real life data set. Therefore, the limit of thresholds can be determined by the users (Breimen, 1984; Quinlan, 1993).

In this study, to determine the significant percentage for the third threshold, few experiments were carried out. Three ranges of percentage were tested which are less than 70%, 70% or more than and more than 80% to discover which range is the best to make the final splitting. Based on the studies that we performed, we strongly recommended the range of 70% or more than as the most suitable percentage to be applied in this research. We presented the average purity from three selected data to prove that the range attained is sufficient to stop the tree from further splitting.

Table 4.1

Percentage Selection for Stopping Rule

Data	Range of percentage	Node	Gini purity index for splitting	Left node	Right node	Average purity
Iris	< 70%	Node 1	0.6521	1.000	0.5016	0.7508
	≥ 70%	-	-	-	-	-
	> 80%	Node 3 Node 4 Node 5	0.8983 0.9571 0.9619	0.8400 1.0000 0.5556	0.9400 0.6250 1.0000	*0.8900 0.8125 0.7778
Pima Indians	< 70%	Node 1	0.6315	0.5997	0.6859	0.6428
	≥ 70%	**Node 3**	**0.7362**	**0.7500**	**0.5900**	***0.6700**
	> 80%	Node 6	0.8233	0.6689	0.6459	0.6574
Bumpus Sparrow	< 70%	Node 1 Node 2	0.5852 0.6049	0.5062 0.5556	0.6800 0.6543	0.5931 0.6050
	≥ 70%	**Node 3**	**0.7714**	**1.0000**	**0.7551**	***0.8776**
	> 80%	Node 5 Node 7	0.8058 0.8000	0.5556 1.0000	1.0000 0.7025	***0.7778** **0.8513**

In Table 4.1, we investigated the percentage of stopping rules by using three ranges (less than 70%, equal or more than 70% and more than 80%) using three famous data sets which are data Iris, Pima Indians and Bumpus Sparrow. From the result, we found that at least 70% is the most reliable cutting point for a node to have its final split. Of course, the higher Gini purity index we gained, the greater homogeneity the node could achieve. However, this may procures a bushy tree and it does not guarantee that the subsequence child nodes could produce a lower overall purity index compare to the previous node. More over, pruning process is required to cut the unfitted sub tree. As mentioned before, this study does not require any post-pruning process since the tree algorithm is taking full protection and accommodation to the data. And, we need to find a significant stopping percentage which could stop the tree from further splitting once it accomplishes the sufficient percentage of homogeneity. In data Iris, coincidentally, there is no Gini index falls in the range of 70% to 80%. But it does not a matter as the node has already achieved a higher percentage of more than 80% in node 3. Node 4 and node 5 are the subsequence nodes of node 3. Although node 4 and node 5 gained a higher Gini purity index for splitting with the value of 0.9571 and 0.9619, the average purity gained is still lower than the one in node 3. For the group of less than 70%, the average purity is only 0.7508 which is considered not sufficient to stop the tree. Thus, we could say that the tree should stop splitting once it has achieved the Gini purity index for equal or more than 70%. In this case, the node 3 sufficiently creates the final terminal nodes. Besides, in Pima Indians data set, node 6 is the child node of node 3. We found that node 3 contains 0.7362 of Gini purity index which split into two child nodes (node 6 and node 7). Since node 7 has gained the minimum number of objects, it stops automatically (as the rule in threshold 2) whereas node 6 is having the potential to split into its subsequence nodes. We computed the Gini purity index for node 6 and we found that node 6 produce even a higher Gini purity index with the splitting value of 0.8233. However, the average purity in its consequence nodes are lower than the one in node 3 (0.67). Therefore, it is clear to prove that 70% or above is sufficient to become the most significant percentage for a node to split into its final nodes rather than taking those in above 80%. We also investigate on the other data called Bumpus Sparrow. In Table 4.1, the group of less than 70% is still unfit to stop due to its low average purity in. Node 5 and node 7 are the child nodes of node 3. Should it be the final split in node 3 or further splitting is needed in node 5 and node 7? Based on the result we gained, node 3 gained the average purity of 0.8776 with its Gini purity index for splitting of 0.7714 (>70%) whereas node 5 and node 7 gained a lower average purity of 0.7778 and 0.8513 respectively although both of them gained a higher Gini purity index for splitting (>80%). Therefore, we can conclude that the percentage of 70% or above is the most significant percentage for a tree to partition into its terminal nodes. In Figure 4.1, we present an example (a path from Pima Indians data set) of this investigation.

Figure 4.1 shows a part of the binary splitting child's nodes from its prior non terminal nodes 1, 3 and 6. To determine which node is the best node to be the final splitting node, Gini purity measurement is carried out. It is a fact that higher Gini purity measurement means that greater purification of the node could produce. In other words, the maximum homogeneity could achieve in its subsequence nodes. However, we have to consider a few criteria such as the size of tree and the accuracy of the tree in particular nodes. Investigating in depth in every node is vital in order to measure the maximization homogeneity the node can produce for the following nodes.

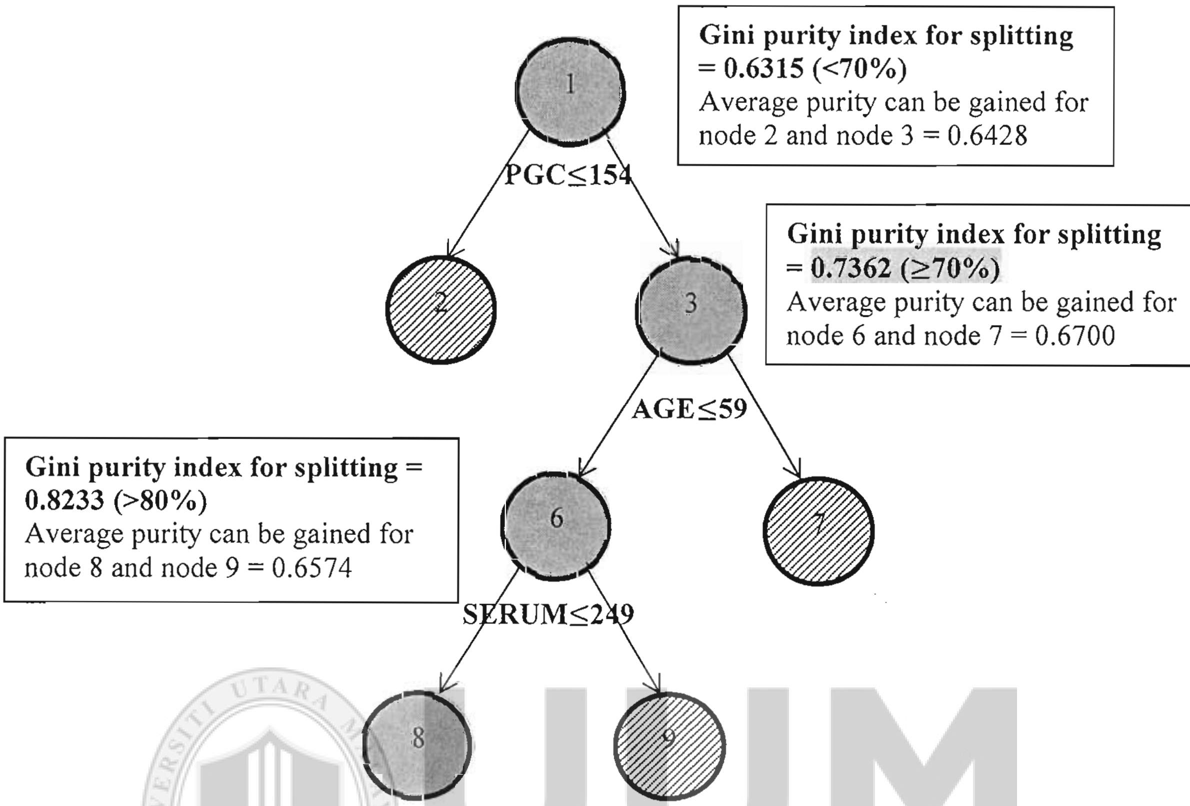


Figure 4.1. Percentage selection for stopping criteria (a path of tree)

We try on the path of node 1, node 3 and node 6 as these paths gone through all the ranges that we set. In node 1, the most potential variable to be chosen is PGC with the splitting point of 154. It successfully divides the observations into the subsequence nodes (node 2 and node 3) with the average purity index of 0.6428 by using Gini purity index of 0.6315 (<70%). Further splitting from node 3 with Gini purity index of 0.7362 (≥70%) to produce node 6 and node 7 which gained the average purity of 0.6700. Then, further split has been carried out from node 6. In this node, SERUM has been selected with the splitting point of 249 as it gained the highest Gini purity index (0.8233) among all the variables. The average purity could be gained for its subsequence nodes (node 8 and node 9) is 0.6574. In this test, we have proven that the node can have its final split once the Gini purity index achieves the percentage of at least 70% (threshold). In this path, we assumed that node 2, node 7 and node 9 are terminal nodes. Only node 1, node 3 and node 6 are inspected for the stopping percentage.

4.3 Case 1: Classification in Breast Tissue Data

The breast tissue data set is a sample of data that explain about breast cancer diagnosis, analysed and reported by some researchers including Jossinet (1996) and Silva, Marques and Jossinet (2000). The measurements in the data are based on Electrical Impedance Spectroscopy (EIS) which are used to measure the complex impedance properties of a material. In medical practices, the EIS measurement of breast tissue can be used as pre screening for cancerous tissue. Therefore, historical data of EIS gives opportunity to researchers to investigate further about the potential patients of breast cancer hence some early pre-cautions can be taken to minimize its implications on the patients.

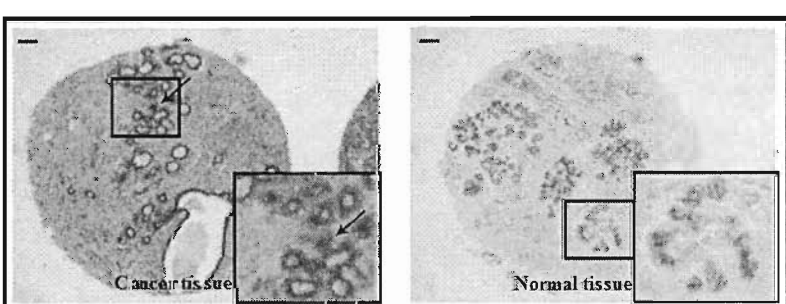


Figure 4.2. Cancer tissue and normal tissue

Breast tissue data set contains nine variables to discriminate 6 classes of tissue namely `car` (carcinoma), `fad` (fibro-adenoma), `mas` (mastopathy), `gla` (glandular), `con` (connective) and `adi` (adipose). The variables that are able to discriminate the groups are based on EIS: impedivity (ohm) at zero frequency (`I0`), phase angle at 500 KHz (`PA500`), high-frequency slope of phase angle (`HFS`), impedance distance between spectral ends (`DA`), area under spectrum (`AREA`), area normalized by DA (`ADA`), maximum of the spectrum (`MaxIP`), distance between `I0` and real part of the maximum frequency point (`DR`), and length of the spectral curve (`P`).

This data set was used by Jossinet (1996) to investigate the variability of impedivity in normal and pathological breast tissue. Overall, the data consists of 106 patients, where 80 of them were used as a training set and the balance is used for assessment. Distributions of patients in each class of tissue are summarized in Table 4.2.

4.3.1 The Statistical Background of Breast Tissue Data

The distribution of patients in each class of tissue is displayed in Table 4.2 and we summarise some statistics about each variable of Breast Tissue in Table 4.3. In this sample, the number of patients in each type of tissue varies across the tissues and all variables have big spread of values as shown by the standard deviation except for `PA500` and `HFS`. Table 4.3 gives some signal of potential outliers in some variables as the recorded skewness value, based on common rule of thumb, is outside the [-2.00, 2.00]. Detail investigation has found that object 64th of variable `AREA` scores

174480.48, quite distinct from the centre point 8142.09 hence could be considered as an outlier.

Table 4.2

Frequency Table of Breast Tissue Data Set

Class of tissue	adi	Car	con	fad	gla	mas	Total
Number of patients	16	16	10	9	14	15	80

Table 4.3

Statistical Description of Breast Tissue Data Set

Variables	Mean	Median	Std. Deviation	Variance	Skewness	Kurtosis
IO	758.72	359.80	749.53	561794.20	1.13	0.26
PA500	0.12	0.11	0.07	0.01	0.96	1.00
HFS	0.12	0.09	0.11	0.011	1.12	1.23
DA	193.36	117.28	203.09	41244.80	1.80	3.72
AREA	8142.09	1814.14	21015.04	4.42E+08	6.59	50.76
ADA	24.34	16.14	25.63	657.10	2.75	11.04
MaxIP	76.64	43.46	86.75	7526.86	2.50	6.44
DR	168.57	93.26	192.43	37029.21	1.88	3.90
P	788.45	431.30	760.48	78331.82	1.26	0.17

Further analysis using graphical presentation as in Figure 4.3 to Figure 4.8 can explain the behaviours recorded in Table 4.2 and Table 4.3. The big spread of data as given by the standard deviation is related to the distinction of classes of tissue which later will be useful for classification purposes as the classes can be identified easily.

Meanwhile, the skewness and kurtosis of `Area` may indicate about the existence of outliers and Figure 4.5 (a) and Figure 4.6 (a) is able to highlight the outlier in the display.

We investigated in details each variable of the breast tissue data to ensure that the data somehow contaminated with outliers. We plotted the distribution of each class of tissues for each variable and spot the separation of the class in a set of displays in Figure 4.4(a), Figure 4.5(a), Figure 4.6(a), Figure 4.7(a) and Figure 4.8(a). Also, we plotted the distribution of the data after the outliers was handled using Winsorize approach in a set of displays from Figure 4.4(b) to Figure 4.8(b). The idea is to spot on the separation between classes of tissue.

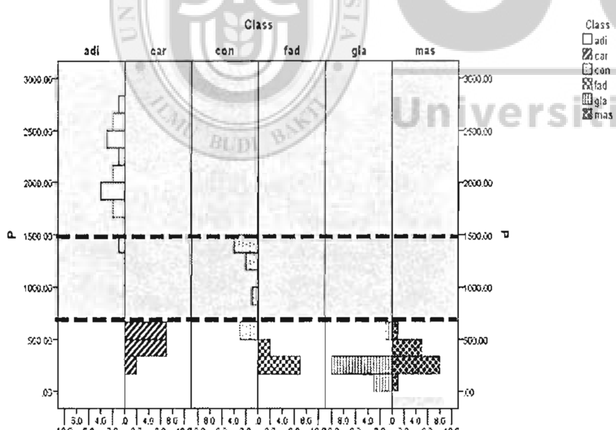


Figure 4.3(a). Original data of variable P

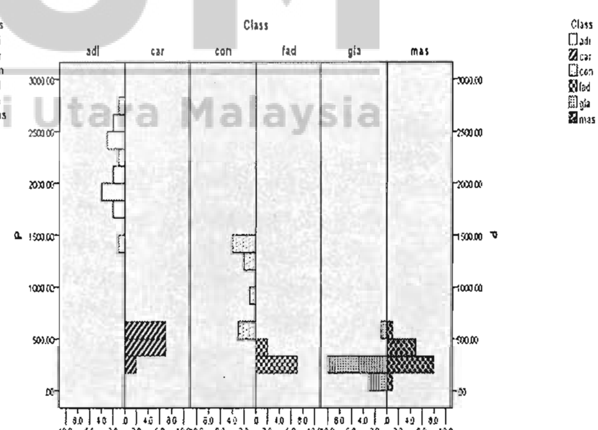


Figure 4.3(b). Winsorize data of variable P

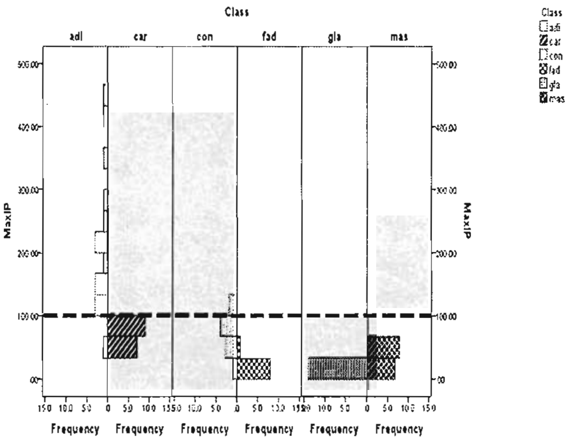


Figure 4.4(a). Original data of variable MaxIP

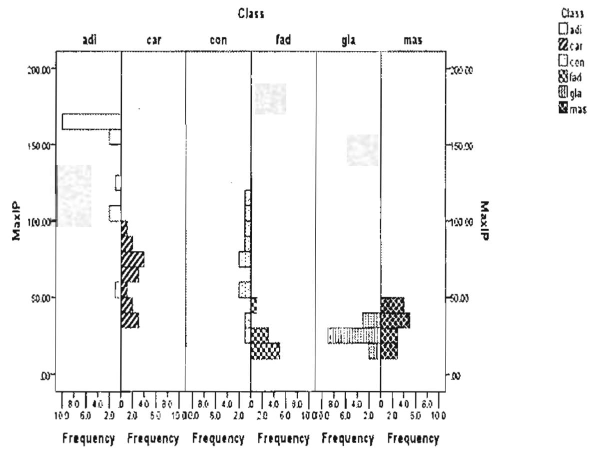


Figure 4.4(b). Winsorize data of variable MaxIP

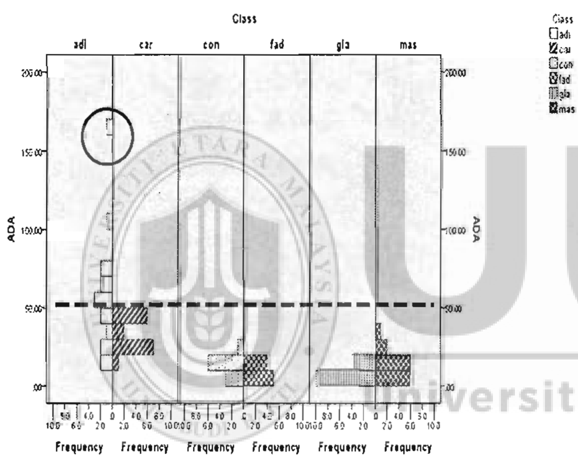


Figure 4.5(a). Original data of variable ADA

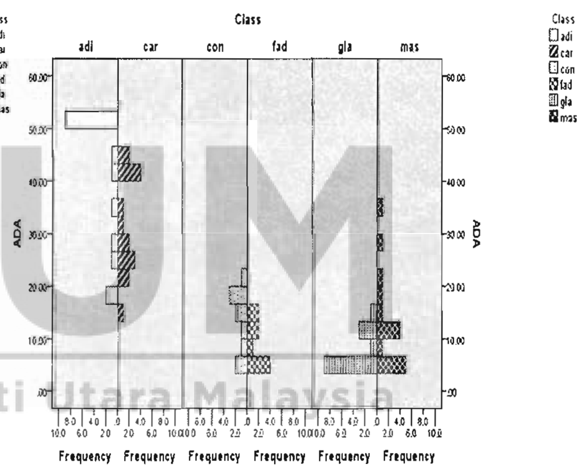


Figure 4.5(b). Winsorize data of variable ADA

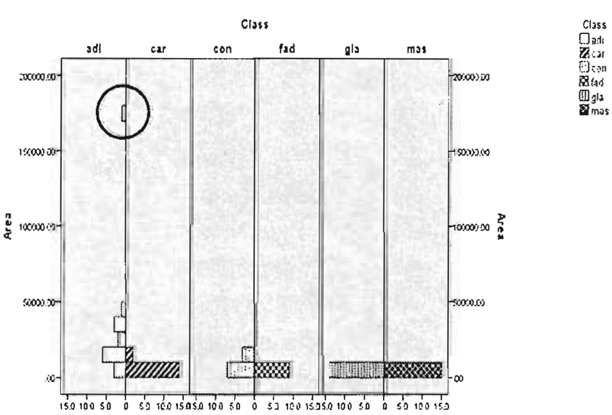


Figure 4.6(a). Original data of variable Area

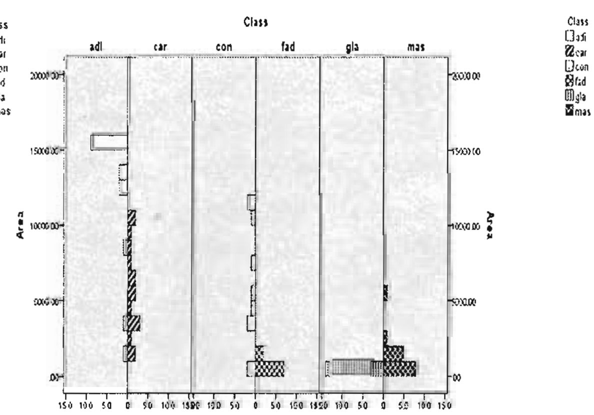


Figure 4.6(b). Winsorize data of variable Area

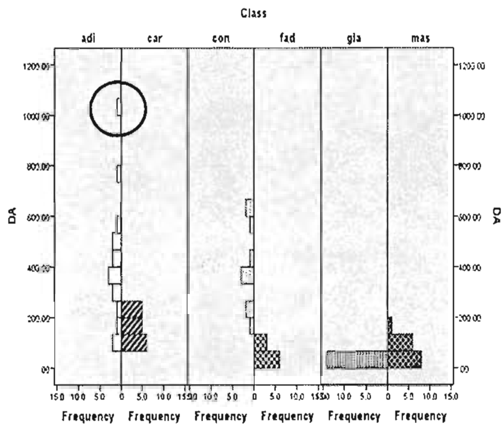


Figure 4.7(a). Original data of variable DA

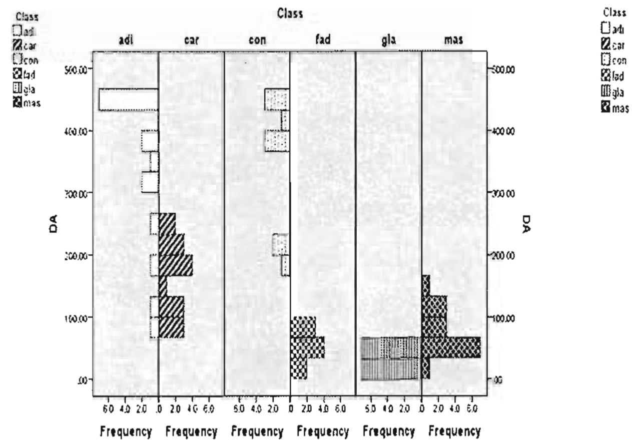


Figure 4.7(b). Winsorize data of variable DA

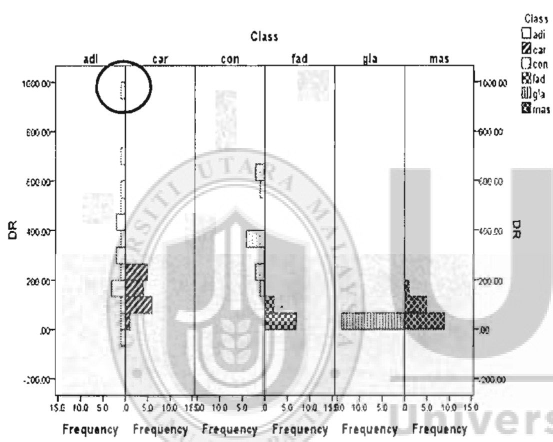


Figure 4.8(a). Original data of variable DR

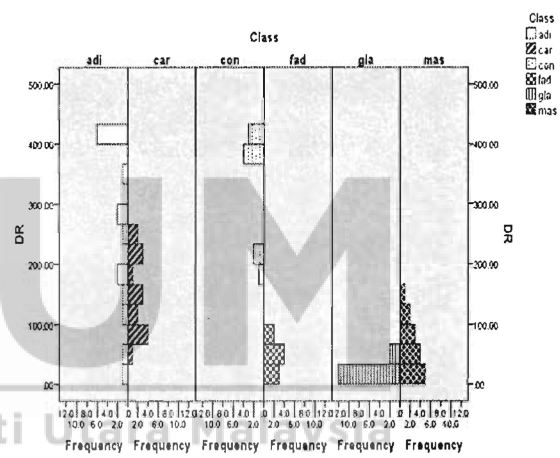


Figure 4.8(b). Winsorize data of variable DR

Based on the plot of original data set (Figure 4.3 to Figure 4.8), we discovered the variables that could be selected are `P`, `IO`, `MaxIP` and `ADA`. These variables may well explain the classes of breast tissue as the redundancy of distribution among the classes is minimal. We mark a single outlier each at Figure 4.5 (a) to Figure 4.8 (a) which may influence the fitted classifier.

We also test for the data normality based on Kolmogorov-Smirnov and Shapiro-Wilk test. According to the result in Table 4.4, we can conclude that all variables are not normally distributed as their p-values are less than 0.05.

Table 4.4

Normality Tests

Variables	**Kolmogorov-Smirnov**[a]			**Shapiro-Wilk**		
	Statistic	**Df**	**Sig.**	**Statistic**	**df**	**Sig.**
PGC	0.058	512	**0.000**	0.969	512	**0.000**
Num_preg	0.156	512	**0.000**	0.907	512	**0.000**
DBP	0.171	512	**0.000**	0.809	512	**0.000**
TRICEP	0.192	512	**0.000**	0.910	512	**0.000**
SERUM	0.250	512	**0.000**	0.707	512	**0.000**
BMI	0.051	512	**0.003**	0.949	512	**0.000**
DPF	0.126	512	**0.000**	0.824	512	**0.000**
AGE	0.150	512	**0.000**	0.872	512	**0.000**

4.3.2 The Construction of Winsorize Tree for Breast Tissue Data

We begin the discussion on breast tissue data set by looking at the earlier stage of tree construction, investigation at the parent node. Using the box plot, 34 outliers have been detected at the node and the number of detected outliers in each variable is tabulated in Table 4.5.

Table 4.5

Outliers in Parent Node

Variables	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
Number of outliers	0	1	2	6	7	4	8	6	0

Next, each variable has gone through a winsorization process at 10% of the both left and right sides of the ordered data. Then, we computed the Winsorize Gini purity index on each variable and chose the variable with the highest score as a splitting variable. For example, PA500 recorded an outlier (see Table 4.6) but the 10% winsorization at the left side of the arranged data of this variable lead to replacement of original values 0.01, 0.02,.., 0.04 with 0.05. The computed Winsorize Gini purity index for values less than 0.05 is 0.2444 and values greater than 0.05 is 0.1924, which give the weighted average at this cutting point as 0.2022. This value indicates that the Gini purity index is still low. The classes are still not clearly separated. Winsorize Gini purity index need to be calculated in order to get the highest weighted average or called Gini purity measurement.

Table 4.6

Example of Winsorize Data and Gini Purity Index for Variable PA500

PA500 (Original)		**PA 500 (Winsorize)**		
0.01	adi	0.05	adi	
0.02	fad	0.05	fad	
0.03	con	0.05	con	
0.04	con	0.05	con	
0.04	con	0.05	con	
0.04	fad	0.05	fad	
0.04	adi	0.05	adi	0.05
0.04	fad	0.05	fad	
0.05	fad	0.05	fad	
0.05	mas	0.05	mas	
0.05	con	0.05	con	
0.05	adi	0.05	adi	
0.06	mas	0.06	mas	
0.06	con	0.06	con	0.06
0.06	gla	0.06	gla	
.	.	.	.	
.	.	.	.	
.	.	.	.	

PA500 (Winsorize)

≤ 0.05					
adi	car	con	fad	gla	mas
3	0	5	4	1	2

> 0.05					
adi	car	con	fad	gla	mas
13	16	5	5	13	13

Gini purity (≤0.05):

$(\frac{3}{15})^2 + (\frac{0}{15})^2 + (\frac{5}{15})^2 + (\frac{4}{15})^2 + (\frac{1}{15})^2 + (\frac{2}{15})^2 = 0.2444$

Gini purity (>0.05):

$(\frac{13}{65})^2 + (\frac{16}{65})^2 + (\frac{5}{65})^2 + (\frac{5}{65})^2 + (\frac{13}{65})^2 + (\frac{13}{65})^2 = 0.1924$

Weighted average:

$$\frac{15}{80}(0.2444) + \frac{65}{80}(0.1924) = 0.2022$$

Table 4.7 summarises computed the highest Gini purity index among eight variables of breast tissue data set at the parent node. Among these variables, P records the highest among the variables with 0.3554 at the splitting point 1428.84. It means P will be used as a variable that split the parent node into left node and right node where the former contains all observations (patients) that score P less or equal than 1428.84, while the right node consists of observations with `P` greater than 1428.84. Based on this cutting point, 63 observations are in the left node and 17 observations are in the right node (node 3).

Table 4.7

Splitting Point in Parent Node

Variable	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
Highest weighted average	0.3467	0.3113	0.2158	0.3031	0.3243	0.3096	0.3317	0.2846	**0.3554**
Location of split	51th	14th	13th	22th	58^{th}	35th	54th	20th	**62th SP: 1428.84**

SP: Splitting point

Once the parent node has been split, the purification of each terminal node is needed to be measured. If the overall Gini purity index in non terminal node achieved more than 0.7, then the node will be considered as leaf or terminal node and no splitting process is necessary to be further carried out. We have discussed earlier that the split of `P` leads to 63 observations in left node (node 2) and 17 observations in right node (node 3). The similar calculation of Winsorize Gini purity index was performed on each variable of each node and ended with purification of node 2 is about 0.2114 and the purification of node 3 is 0.8892 (see Figure 4.9). Between these two nodes, the node 3 is almost pure with 0.8892 (achieved one of the threshold) and detail of investigation has found out that the node contains 16 observations from the group `adi` and only 1 observation from `con` (see Table 4.8). Since the threshold is met, node 3 is considered as terminal node.

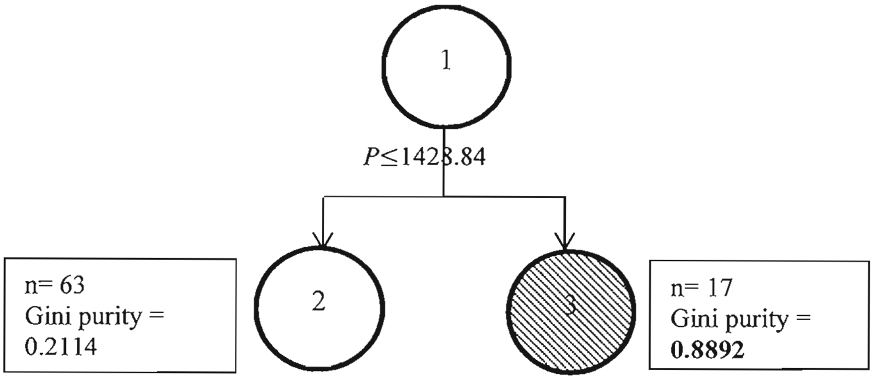


Figure 4.9. Splitting of parent node

Table 4.8

Number of Observations in Node 2 and Node 3

Node \ Group	adi	Car	Con	fad	gla	mas
Node 2	0	16	9	9	14	15
Node 3	16	0	1	0	0	0

Node 2 scores low Gini purity index as the node consists of complexity of group memberships. Group domination is not clear at this stage hence more splitting processes need to be considered. In node 2, the process of splitting as in node 1 is repeated where outliers must be inspected in all variables by using the original data available at node 2. There are 29 outliers are found from 63 observations. Again, Winsorize method is applied to neutralised the heavily tails before performing Gini measurement. Table 4.9 showed the result of Gini purity index in node 2.

Table 4.9

Splitting Point in Node 2

Variable	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
Highest weighted average	0.3657	**0.3930**	0.2754	0.3601	0.3674	0.3838	0.3495	0.3538	0.3407
Location of split	45th	**14th** **SP: 0.18**	14th	24th	32th	44th	37th	23th	31th

At node 2 (see Figure 4.10), the computed Gini indexes showed that PA500 is the best splitting variable with the splitting point 0.18. By using this information, we split the observations accordingly which led to 48 observations at node 4 (less or equal to 0.18) and 15 observations to node 5 (more than 0.18). Careful assessment of both nodes 4 and 5 found out that node 5 achieved purity score 0.8711 (more than 0.7) hence it was flag as pure and no further splitting process is necessary.

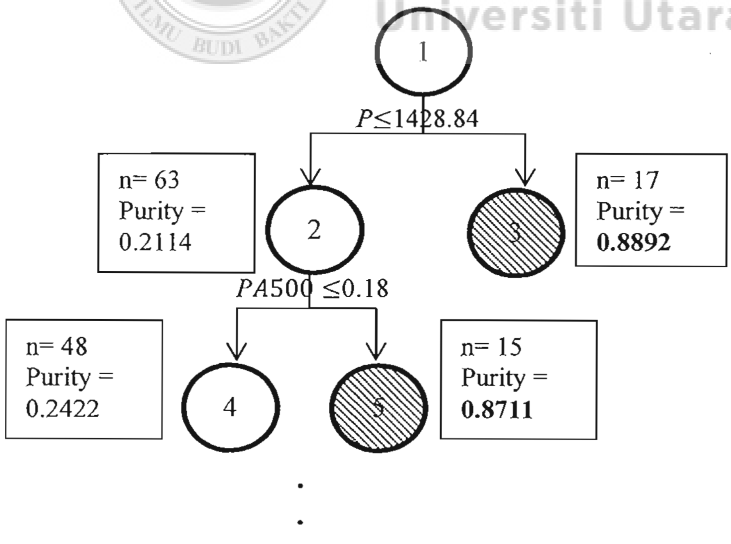


Figure 4.10. Child nodes from node 2

Table 4.10

Number of Observations in Node 4 and Node 5

Group / Node	adi	car	Con	fad	gla	mas
Node 4	0	2	9	9	14	14
Node 5	0	14	0	0	0	1

Table 4.10 shows the distribution of observations at node 4 and node 5. Node 4 contains observations in all classes of tissue except `adi`, while node 5 is dominated by observations from `car`. This distribution explains well the purity index recorded by node 4 and node 5 as previously discussed.

In the identification process of outliers, winsorization of values on the detected variables with outliers and determination of best split were performed at each node until each node meets its terminal based on one of the three thresholds. Once each node has met the final terminal, then we obtained a full constructed tree which summarises the whole process of classification.

In comparison to the proposed Winsorize tree, the traditional tree may isolate real outliers in any terminal node. However, keeping the outliers throughout the process of tree construction may increase time for analysis purposes and produce a tree with many branches. Such phenomenon called bushy tree is not helpful in assisting practitioners to predict the group of future observations. Therefore, a pruning process can be considered so that an acceptable size of tree can be structured. But, as we have discussed in Chapter 2, the pruning process is merely for an expert rather than practitioners.

We constructed both traditional tree and pruned tree to be compared to the Winsorize tree. The Winsorize tree is as depicted in Figure 4.11, traditional tree can be seen in Figure 4.12 and the pruned tree in Figure 4.13. By using naked eyes, we can detect small differences among these three trees. The Winsorize tree shows great branches on the left side and it uses variable `P` as a splitting variable in the parent node. Meanwhile, both traditional tree and pruned tree show similar structure with variable `I0` as a splitting variable. As the matter of fact, the pruned tree has the similar structure as the traditional tree but with fewer leaves. The next section will discuss about the overall assessment of these trees.

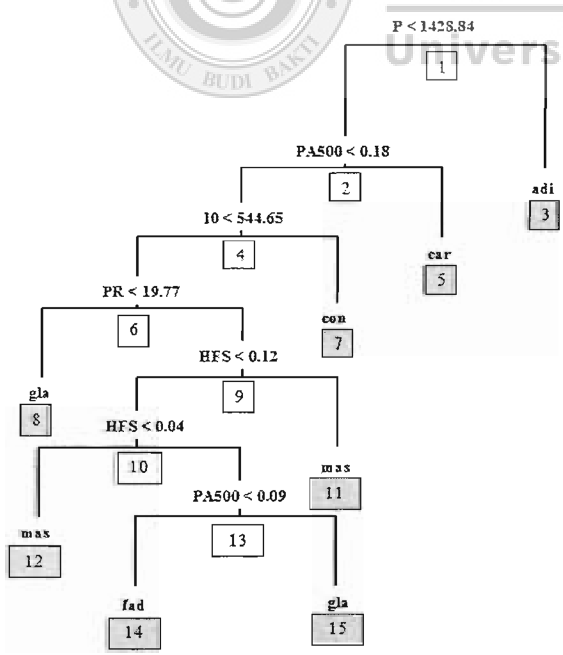


Figure 4.11. Winsorize tree of Breast Tissue

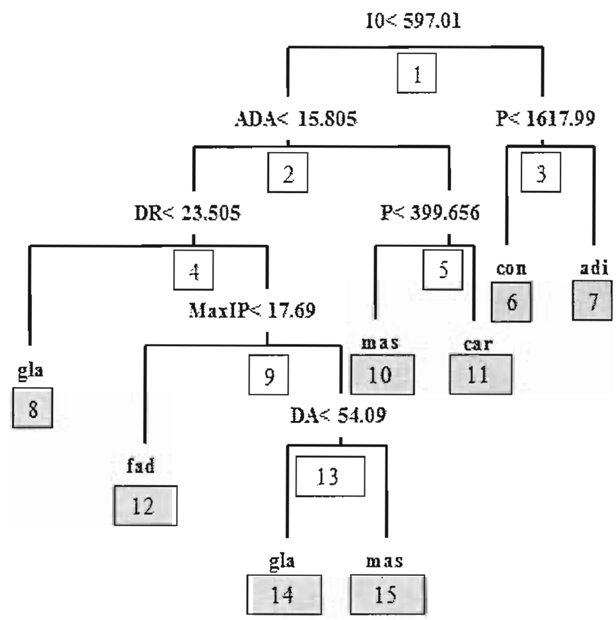


Figure 4.12. Traditional tree of Breast Tissue

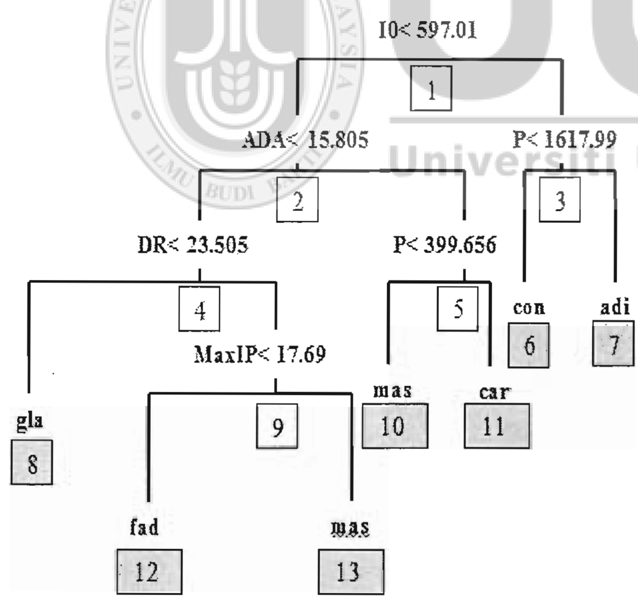


Figure 4.13: Pruned tree of Breast Tissue

4.3.3 The Evaluation of Winsorize Tree for Breast Tissue Data

We evaluated the constructed tree based on several criteria: (i) structure of tree, (ii) error rate and (iii) number of outliers detected. The summaries of each constructed tree are given in Table 4.11.

Table 4.11

Comparison between Traditional Tree, Pruned Tree and Winsorize Tree

BREAST TISSUE:	Traditional Tree	Pruned Tree	Winsorize Tree
i. Number of splitting	7	6	7
ii. Number of leaves	8	7	8
iii. Number of variable used	6	5	5
iv. Name of variables used	1. P 2. I0 3. ADA 4. DR 5. MaxIP 6. DA	1. P 2. I0 3. ADA 4. DR 5. MaxIP	1. P 2. PA 500 3. I0 4. PR 5. HFS
v. Error rate	0.3846	0.4231	***0.2308**
vi. Outliers detected:			
a. First node	-	-	34
b. Second node	-	-	29
c. Forth node	-	-	41
d. Sixth node	-	-	10
e. Ninth node	-	-	2
f. Tenth node	-		1
g. Thirtieth node	-		0

In term of structure, pruned tree uses the fewest number of leaves and split. Although the proposed Winsorized tree has similar tree structure to the traditional tree and number of variables, the former uses different variables for classifying the breast cancer patients except for P (length of the spectral curve) and I0 (impedivity (ohm) at zero frequency). These results indicate that both variables are important in explaining the differences between the six classes of tissue. Many researchers believed that the tree itself can isolate the outliers without affecting the classification, but our results show that ignoring the outliers can produced an inaccurate tree. Many outliers have been detected in different nodes. The suspicious values affected the Gini index measurement and the cutting points. As a consequences, many insensible branches could be produced which lead to bias result.

The result has proven that Winsorize tree produced the lowest error (0.2308) compared to the traditional tree and the pruned tree. We believe that such result can be explained by the process of detecting and handling the outlier, skewness of data caused by outliers has been solved. Due to the impurity of the data are reduced, the Gini index measurement becomes more accurate. This means real sensible variables with best split will be selected to construct tree. In short, Winsorize tree produces a comparative tree with no pruning process and low error rate. In addition, all outliers in all nodes are well treated and only true attributes are selected as the splitting attribute.

4.4 Case 2: Classification in Egyptian Skull Data

Four measurements were made of male Egyptian skull from five different time period ranging from 4000B.C to 150 A.D. The changes of skull sizes were recorded between the time periods. The researchers theorize that the change in skull size is due to the interbreeding of the Egyptians with immigrant population over the years. (Egyptian Skull Development. (n.d.). *StatLib and Story Library.* Retrieved June, 2014, from http://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html). Egyptian skull data set contains 150 number of cases which 113 cases are used as training set and the rest are used as test set. Four measurements of male Egyptian skull which are `mb` (maximal breadth of skull), `bh` (basibregmatic height of skull), `bl` (basialveolar length of skull) and `nh` (nasal height of skull) from 5 different time periods (negative = BC, positive = AD) (`epoch`) are recorded. Tree is used to categorise the skull size over the time period.

4.4.1 The Statistical Background of Egyptian Skull Data

The distribution of 113 skulls of the training set based on five time periods is tabulated in Table 4.12. The training set consists of the similar number of sample of skull across the period of time. Meanwhile, Table 4.13 summarises some descriptive statistics in order to give an overview about the behavior of each measured variables namely `mb`, `bh`, `bl` and `nh`. The estimated mean and median for all variables seem similar hence none of the variables may consist outliers. The values of kurtosis and the values of skewness do not indicate the sign of having outliers. Therefore, we may conclude that the empirical evidences of Egyptian skull data set are free from outliers and could have symmetry distributions.

Table 4.12

Frequency Table of Eqyptian Skull Data Set

Epoch	**C1850BC**	**C200BC**	**C3300BC**	**C4000BC**	**cAD150**	**Total**
Frequency	21	23	26	22	21	113

Table 4.13

Statistical Description of Egyptian Skull Data Set

Variables	**Mean**	**Median**	**Std. Deviation**	**Variance**	**Skewness**	**Kurtosis**
Mb	133.97	134.00	4.82	23.22	-0.01	0.51
Bh	132.65	133.00	5.04	25.41	-0.17	0.19
Bl	96.49	96.00	5.16	26.57	-0.09	0.06
Nh	50.89	51.00	3.16	9.99	0.11	-0.18

Further investigation was carried out on each variable according to the period of time. Figure 4.14 and Figure 4.15 display the bar charts of each class (period of time) against two selected variables, nh and bl. Both displays attempt to highlight separation between classes and the sign of outliers in the variables.

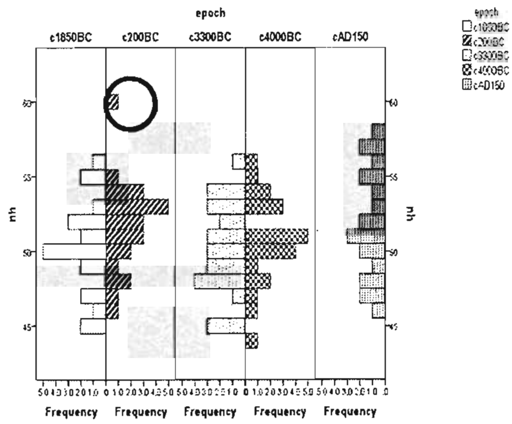


Figure 4.14(a). Original data of variable nh

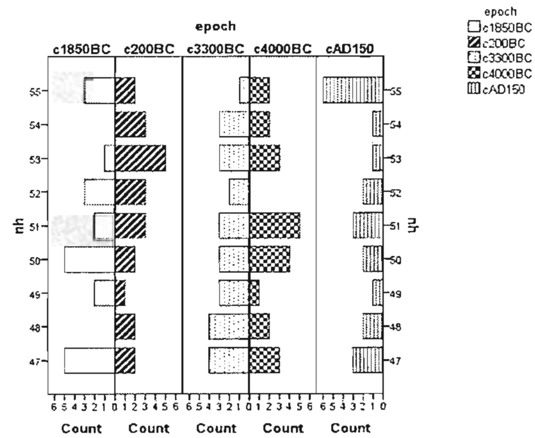


Figure 4.14(b). Winsorize data of variable nh

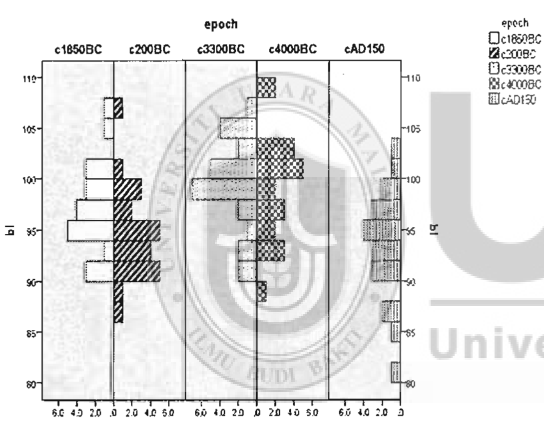


Figure 4.15(a). Original data of variable bl

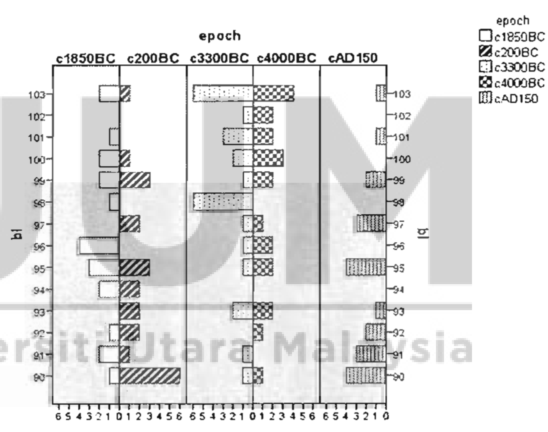


Figure 4.15(b). Winsorize data of variable bl

The spread of observation as in circle in Figure 4.14 (a) can be regarded as potential outlier. Egyptian skull data set is a complicated classification case as most of the classes are greatly redundant onto each other. None of the display in Figure 4.14 and Figure 4.15 show a clear cut between classes. Another investigation on the other two variables, `mb` and `bh`, based on scatterplot as in Figure 4.16 discover the swamp of observations hence separation lines between classes are hardly to be spotted too. Few potential outliers can be observed as in circles. However, when Winsorize method is performed, the extreme values are replaced by where the floor and the ceiling of the observations are dragged to the range from 126 to 138 as in Figure 4.16 (b).

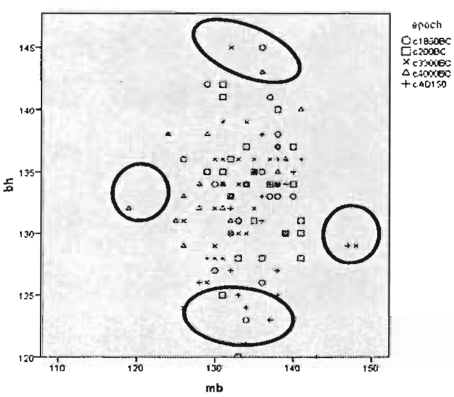


Figure 4.16(a). Scatterplot of bh against mb

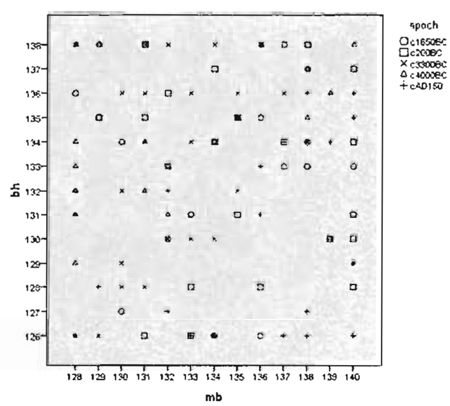


Figure 4.16(b). Scatterplot of bh against mb using Winsorize method

Table 4.14

Normality Tests

Variables	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
Num_preg	0.16	512	**0.00**	0.91	512	**0.00**
PGC	0.06	512	**0.00**	0.97	512	**0.00**
DBP	0.17	512	**0.00**	0.81	512	**0.00**
TRICEP	0.19	512	**0.00**	0.91	512	**0.00**
SERUM	0.25	512	**0.00**	0.71	512	**0.00**
BMI	0.51	512	**0.00**	0.95	512	**0.00**
DPF	0.13	512	**0.00**	0.82	512	**0.00**
AGE	0.15	512	**0.00**	0.87	512	**0.00**

According to the result of normality test in Table 4.14, both tests (Kolmogorov-Smirnov and Shapiro-Wilk) show that all the variables are not normal distributed as the p-value is less than 0.05.

4.4.2 The Construction of Winsorize Tree for Egyptian Skull Data

The boxplot is capable to identify some outliers from each variable of the skull data.

Table 4.15

Outlier in Parent Node

Variable	mb	bh	bl	nh
Number of outliers	1	2	1	1

All these suspicious values have been Winsorize at 10%, followed by the computation of the Gini purity index to determine the most potential variable to be used as a split variable in the parent node. Among these variables, bl gives the highest weighted average hence it is chosen in the first split with the spitting value, 96. The table of Gini purity index is showed below.

Table 4.16

Splitting Point in Parent Node

Variable	Mb	bh	bl	nh
Highest weighted average	0.2299	0.2291	0.2408	0.2143
Location of split	8^{th}	2th	8th SP: 96	8th

For the splitting process, those observations with the *bl* less than or equal to 96 will be assigned to the left node, t_l, and the remaining observations will be assigned to the right node, t_r. There are 57 observations and 56 observations of the original data are split into left (node 2) and right node (node 3) respectively as shown in Figure 4.17.

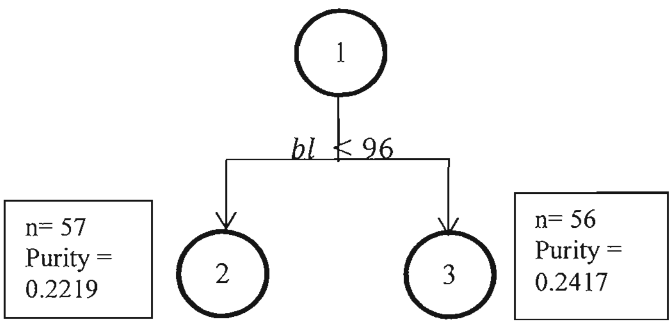


Figure 4.17. Child nodes from node 1

Table 4.17

Number of Observations in Node 2 and Node 3

Node \ Group	C1850BC	C200BC	C3300BC	C4000BC	cAD150
Node 2	13	16	6	8	14
Node 3	8	7	20	14	7

Based on Figure 4.17, total overall purity measurement in each node 2 and node 3 are considered low, where both are below than 0.25. The purity rate has not reached the target of threshold, 0.70 or the minimum value of n. This phenomenon is due to the complexity of group causing the data are hardly to be cut. Therefore, further splitting is needed in order to gain a purer node.

In the second node, the process was repeated where outliers were inspected again in all variables using the original data set. In the left node (node 2), 3 outliers have been detected from each variable except bh (Table 4.18). In contra, the right node (node 3) contains 5 outliers (Table 4.19). Again, Winsorize method is applied in both nodes to neutralise the heavy tails before performing Gini purity index computation.

Table 4.18

Outliers in Node 2

Variable	mb	bh	bl	nh
Number of outliers	1	0	1	1

Table 4.19

Outliers in Node 3

Variable	mb	bh	bl	nh
Number of outliers	1	2	2	0

Table 4.20

Gini Index of Winsorize Tree in Node 2

Variable	Mb	bh	bl	nh
Highest weighted average	0.2580	0.2689	0.2611	0.2420
Location of split	11^{th}	9th SP:129	8th	9^{th}

Table 4.21

Gini Index of Winsorize Tree in Node 3

Variable	mb	Bh	bl	nh
Highest weighted average	0.2688	0.2607	0.2696	0.2741
Location of split	1st	1st	1st	5th SP:49

According to Table 4.20 and Table 4.21, Gini purity index shows that bh is the best splitting variable with the splitting point 129 in node 2 (0.2689) while nh is the best splitting variable with the splitting point 49 in node 3 (0.2741). In node 2, 18 observations are moved to node 4 and the remaining are assigned to node 5; 17 observations and 39 observations are move to the node 6 and node 7 respectively. The details of splitting process in second level are displayed in Figure 4.18.

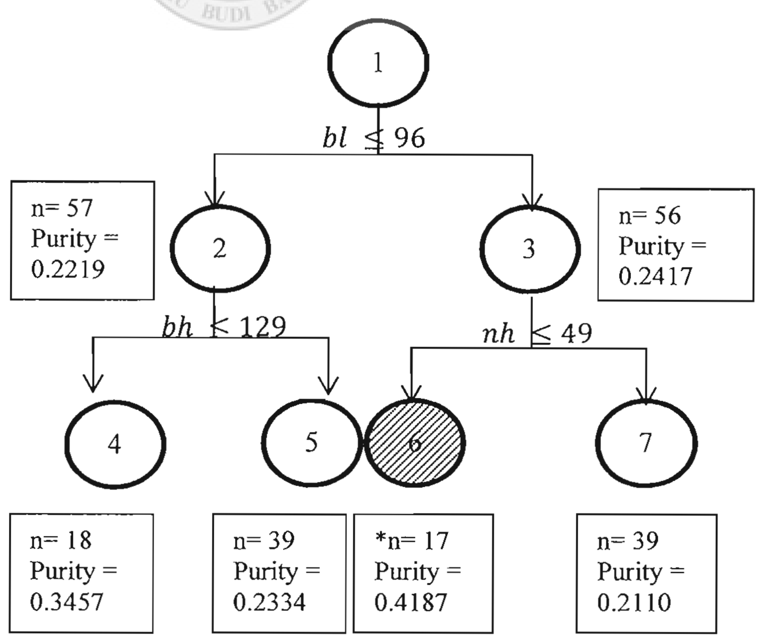


Figure 4.18. Child nodes from node 2 and node 3

The splitting process at the second level as depicted by Figure 4.18 and the summary of number of observations in Table 4.22 shows that node 6 has reached it terminal node. The distribution of classes in node 6 reveals that the number observations in this node is less than or equal to the set up threshold, $n_{min} = 15\% \times n$ which is 17. Therefore, node 6 is the terminal node. In contrary, further splits are conducted on node 4, node 5 and node 7. The process is repeated recursively until the nodes achieve one of the three thresholds.

Table 4.22

Number of Observations in Node 4, Node 5, Node 6 and Node 7

Group / Node	C1850BC	C200BC	C3300BC	C4000BC	cAD150
Node 4	1	5	2	1	9
Node 5	12	11	4	7	5
Node 6	2	0	10	4	1
Node 7	6	7	10	10	6

The final structure of the Winsorize tree on the Egyptian skull data set is displayed in Figure 4.19. Also, Figure 4.20 and Figure 4.21 is a traditional tree and a pruned tree constructed on the Egyptian skull data set. Discussions on the comparisons among the three trees are given in the next subsection.

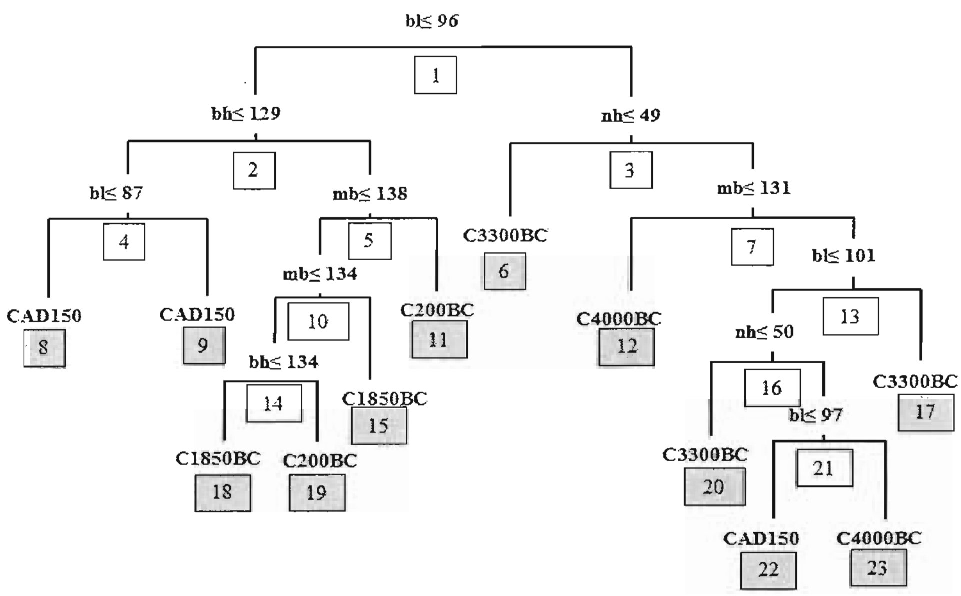


Figure 4.19. Winsorize tree of Egyptian Skull

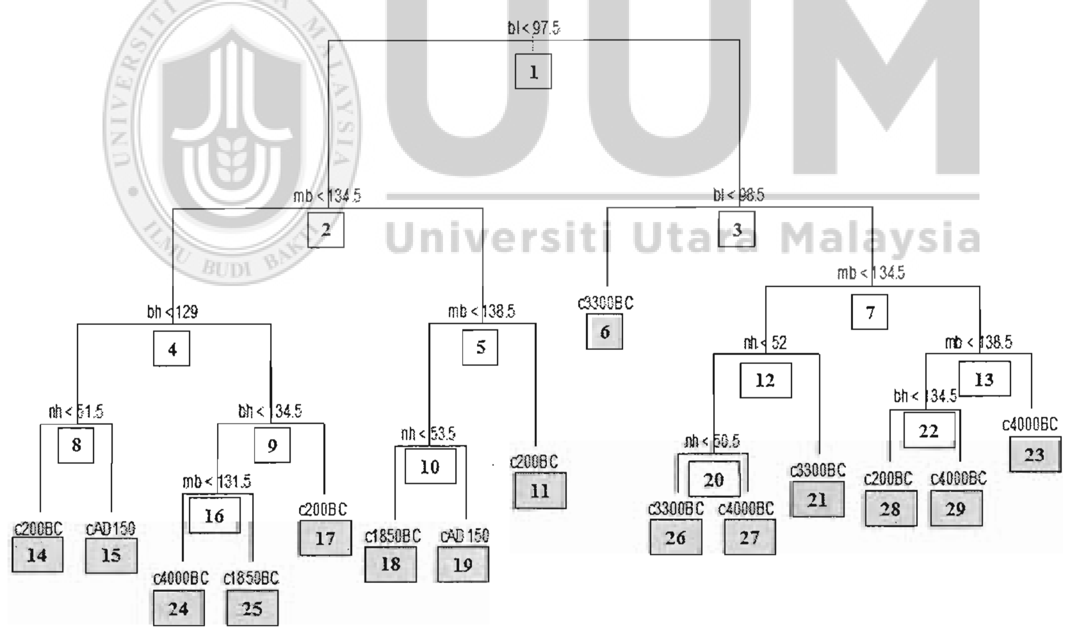


Figure 4.20. Traditional tree of Egyptian Skull

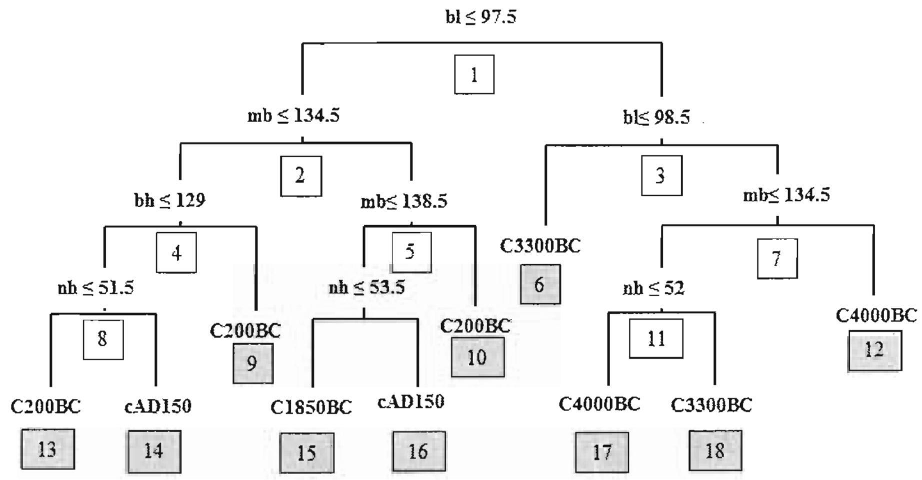


Figure 4.21. Pruned tree of Egyptian Skull

4.4.3 The Evaluation of Winsorize Tree for Egyptian Skull Data

In this section, detail discussion is carried out to compare on these three types of constructed tree. Some performances of interest of these trees are displayed in Table 4.23.

Table 4.23

Comparison between Traditional Tree, Pruned Tree and Winsorize Tree

EGYPTIAN SKULL:	**Traditional Tree**	**Pruned Tree**	**Winsorize Tree**
i. Number of splitting	14	9	11
ii. Number of leaves	15	10	12
iii. Number of variable use	4	4	4
iv. Name of variable used	1. bl 2. mb 3. nh 4. bh	1.bl 2.mb 3.nh 4.bh	1. bl 2. mb 3. nh 4. bh
v. Error rate	**0.8108**	***0.7568**	***0.7568**

EGYPTIAN SKULL:	Traditional Tree	Pruned Tree	Winsorize Tree
vi. Extreme value detected:			
a. First node	-	-	5
b. Second node	-	-	3
c. Third node	-	-	5
d. Forth node	-	-	1
e. Fifth node	-	-	3
f. Seventh node	-	-	6
g. Tenth node	-	-	3
h. Thirteenth node	-	-	0
i. fourteenth node	-	-	3
j. sixteenth node	-	-	2
k. twenty-first node	-	-	2

All constructed trees used all the measured variables (`bh, bl, mb` & `nh`), which tell us that all these four variables may discriminate the skull size according to period of time. Despite of this similar behavior, traditional tree records the highest error rate which is 0.8108. Besides, it has a bushy structure with greatest number of leaves and splits compared to the pruned tree and Winsorize tree. In contrast, pruned tree produces the smallest tree with nine splits. Winsorize tree produces medium size tree with error error rate 0.7568 which is at the same performance to the pruned tree. Although Winsorize tree has bigger size of tree compared to pruned tree, it might be the most reliable tree as all the outliers are successfully been detected and handled. Moreover, no pruning process is required as the tree stopped splitting when one of our thresholds is met.

4.5 Case 3: Classification in Pima Indians Data

The Pima Indians diabetes database was donated by Vincent Sigillito in year 1990 from a population of Phoenix, Arizano, USA. The data set contains the collection of medical diagnosis report of 768 observations and 9 variables with two dependent variables on the status of diabetes, either Positive (P) or Negative (N) of getting diabetes. There are 500 patients from the Negative group and the remaining are from Positive group being tested with positive for diabetes in the 2 hours post-load plasma glucose was at least 200mg/dl. In particular, the patients are female at least 21 years old from Pima Indians heritage (Smith, Everhart, Dickson, Knowler and Johannes, 1988). The variables used for distinguishing those suffer with diabetes are as below:

1. Number of times pregnant [`Num_preg`]
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test [`PGC`]
3. Diastolic blood pressure (mm Hg) [`DBP`]
4. Triceps skin fold thickness (mm) [`Tricep`]
5. 2-Hour serum insulin (mu U/ml) [`SERUM`]
6. Body mass index (weight in kg/(height in m)^2) [`BMI`]
7. Diabetes pedigree function [`DPF`]
8. Age (years) [`Age`]
9. Class variable (0 or 1) [`Class P or N`]

4.5.1 The Statistical Background of Pima Indians Data

The Pima Indians data set consists of patients diagnosed positive or negative with diabetes. Table 4.24 displays the distribution of the sample of patients of these two groups.

Table 4.24

Frequency Table of Pima Indians Data Set

Class variable	**Positive**	**Negative**	**Total**
Frequency	185	327	512

Table 4.25

Statistical Description of Pima Indians Data Set

Variables	**Mean**	**Median**	**Std. Deviation**	**Variance**	**Skewness**	**Kurtosis**
Num_preg	3.82	3.00	3.39	11.28	0.93	0.40
PGC	120.92	117.00	32.85	1079.04	0.00	0.88
DBP	68.83	70.00	19.30	372.28	-1.89	5.38
TRICEP	20.65	23.00	15.47	244.84	-0.03	-1.12
SERUM	80.09	23.00	118.88	14130.57	2.42	8.11
BMI	31.97	36.00	8.15	66.38	-0.37	3.23
DPF	0.49	32.00	0.34	0.12	2.06	6.38
AGE	33.14	0.38	11.65	135.82	1.20	0.92

Some basic statistics measurements are tabulated in Table 4.25. The value of standard deviation is high especially in PGC and SERUM indicates that the data points are wide spread from the data. And, the distribution of data is skewed, probably due to the occurrence of outliers. Besides, these variables have high degree of peakness (called leptokurtic distribution). To confirm the distribution of these variables, further inspection and investigation need to be carried out.

We plotted the graphs which can help us to over view the distribution of the data. Suspicious values also can be detected. Here, we displayed three population pyramid graphs to show that the present of outliers and perform Winsorize towards the outliers. The graphs are shown in Figure 4.22 to Figure 4.25.

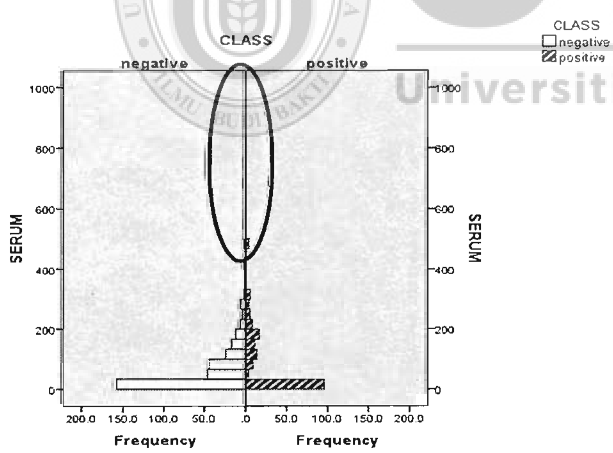


Figure 4.22 (a). Original data of variable SERUM

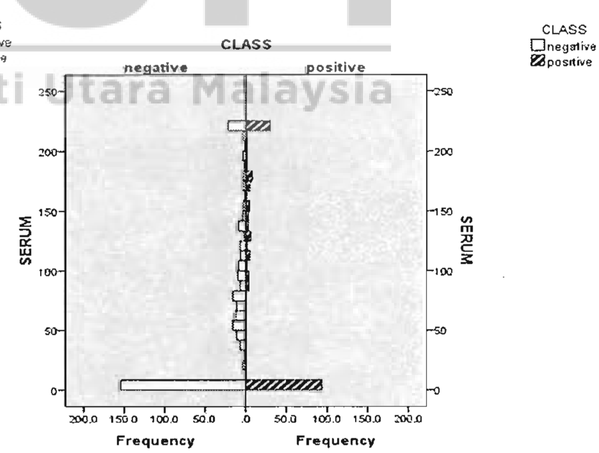


Figure 4.22 (b). Winsorize data of variable SERUM

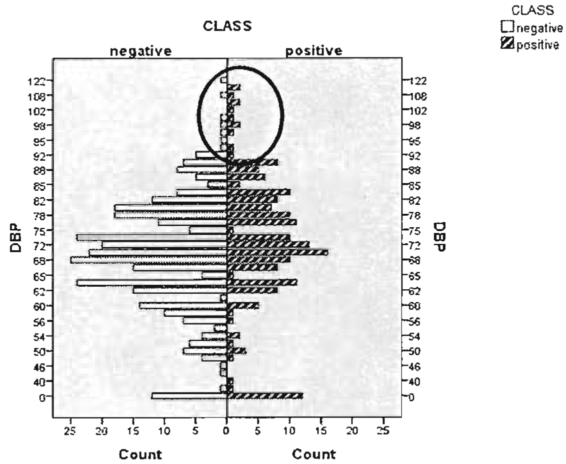


Figure 4.23 (a). Original data of variable DBP

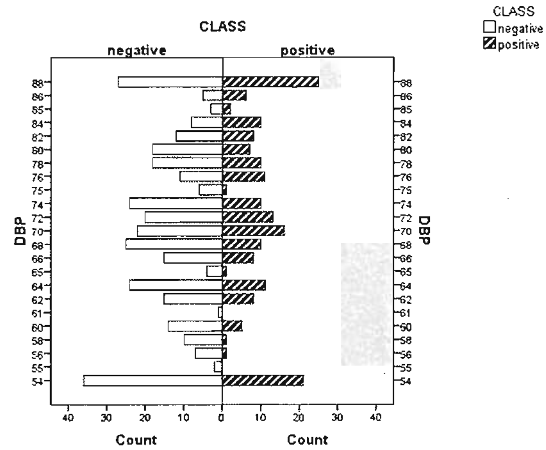


Figure 4.23 (b). Winsorize data of variable DBP

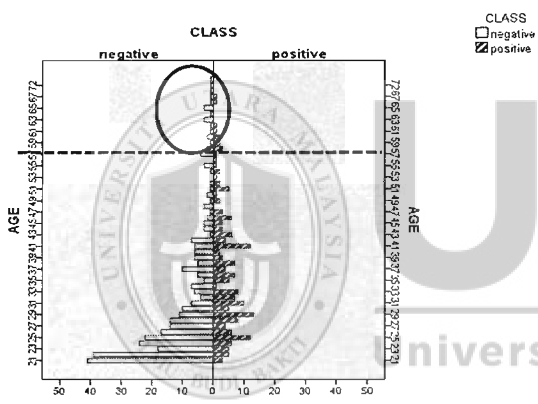


Figure 4.24 (a). Original data of variable AGE

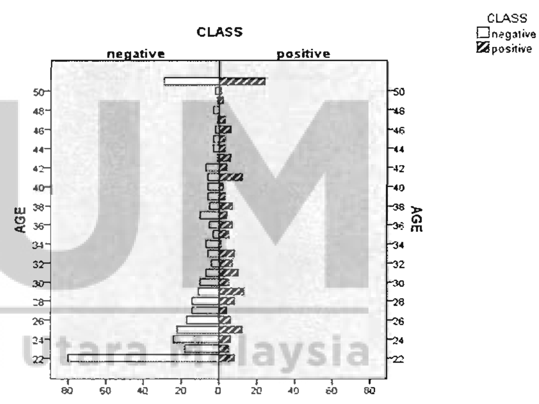


Figure 4.24 (b). Winsorize data of variable AGE

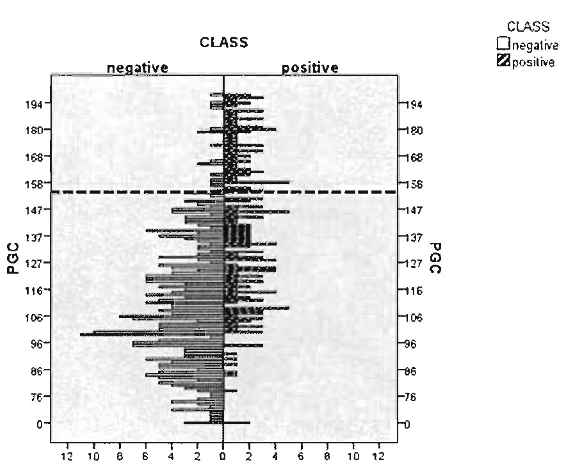


Figure 4.25 (a). Original data of variable PGC

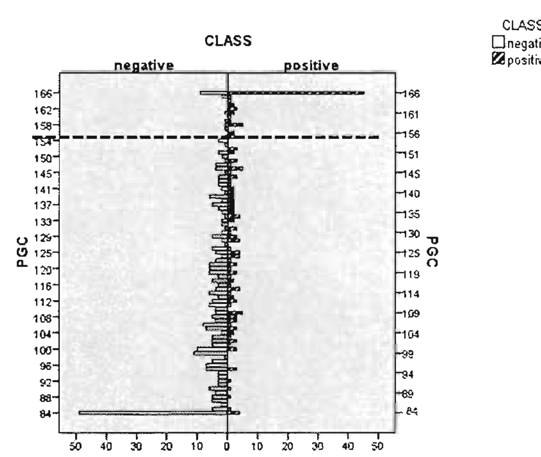


Figure 4.25 (b). Winsorize data of variable PGC

Based on the Figure 4.22 (a), obviously, SERUM has a very long tail. The mark of circle highlights the possibility of outliers, which are then Winsorize to norm as shown in Figure 4.22 (b). Since all the classes are overlapping onto each other, the clear cutting point is hardly to be identified. Roughly, the potential point is at the splitting point of 154 in variable PGC.

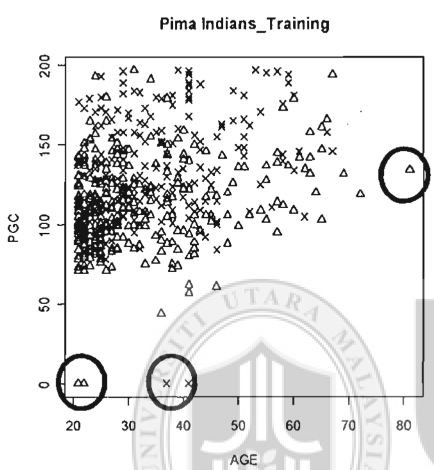


Figure 4.26(a). Scatterplot of original Pima Indians training data set

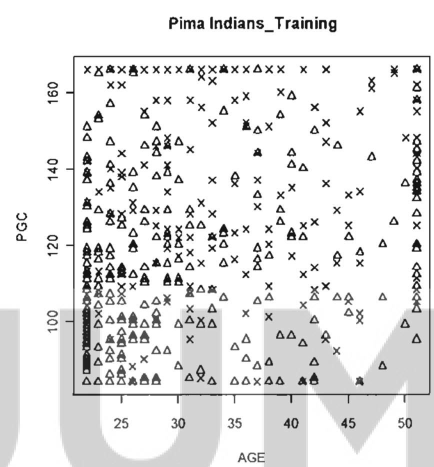


Figure 4.26(b). Scatterplot of Winsorize Pima Indians training data set

Due to its complexity, variable PGC against variable AGE is drawn using scatter plot (see Figure 4.26). The figure shows that both classes are swamped together especially when AGE is less than 38 and PGC is less than 150. We highlight the potential outliers in circles. Therefore, Winsorize the data is necessary to reduce the effect of outliers while constructing tree.

4.5.2 The Construction of Winsorize Tree for Pima Indians Data

Based on boxplot in Figure 4.27, 109 outliers have been detected in Pima Indians data set where the greatest number of outliers were in DBP with 31 outliers and followed by SERUM with 25 outliers. Details on the recorded number of outliers for each variable are given in Table 4.26.

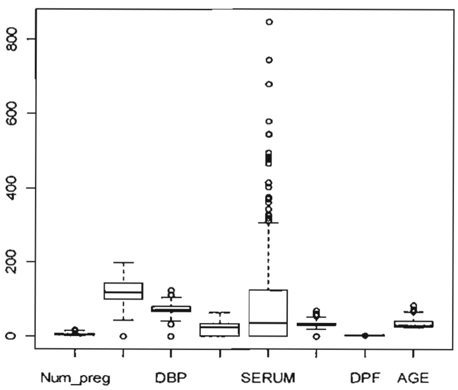


Figure 4.27. Outliers detection using boxplot

Table 4.26

Outliers in Parent Node

Variable	**Num_preg**	**PGC**	**DBP**	**TRICEP**	**SERUM**	**BMI**	**DPF**	**AGE**
Number of Outliers	4	5	31	0	25	15	19	10

The process of handling outliers is similar to the one discussed in subsection 4.1.2 where those identified outliers have to be Winsorize prior to the computation of the Gini purity measurement.

Table 4.27

Splitting Point in Parent Node

Variable	Num_preg	PGC	DBP	TRICEP	SERUM	BMI	DPF	AGE
Highest weighted average	0.5603	0.6135	0.5462	0.5494	0.5608	0.5877	0.5534	0.5814
Location of split	7th	69th SP: 154	19th	25th	68th	49th	263th	6th

The computed Gini purity measurement as in Table 4.27 indicates that PGC recorded the highest weighted average with the value of 0.6135 at the splitting point 154. It means that this variable is the best variable to be split at the parent node. Following this, 430 objects are assigned to the left node, which consist of 311 patients from negative class and 119 patients are from positive class. In contrary, 82 patients are assigned to the right node which consists of 16 patients from negative class and 66 patients from positive class. The structure of the first split and the total number of patients are shown in Figure 4.28 and Table 4.28.

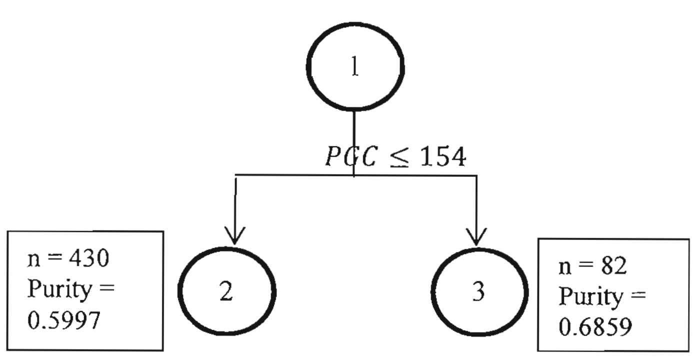


Figure 4.28. Child nodes from parent node

Table 4.28

Number of Patients in Node 2 and Node 3

Group / Node	N	P
Node 2	311	119
Node 3	16	66

In node 2, the overall Gini purity is approximately 0.6. Node 3 is purer as it gains a higher Gini index which is 0.6859. Since both nodes are still below the threshold (Gini purity index more than 0.7), further splitting process is needed.

In node 2 and node 3, the process of identifying outliers is repeated as in node 1. There are 109 outliers and 14 outliers found in node 2 and node 3 respectively. In node 2, DBP consists the highest number of outliers which is 28 while TRICEP has no outlier at all. No outlier is detected in variable PGC, TRICEP and AGE in node 3. Details can refer to Table 4.29 and Table 4.30.

Table 4.29

Number of Outliers in Node 2

Variable	Num_preg	PGC	DBP	TRICEP	SERUM	BMI	DPF	AGE
Number of Outliers	15	6	28	0	19	12	14	15

Table 4.30

Number of Outliers in Node 3

Variable	Num_preg	PGC	DBP	TRICEP	SERUM	BMI	DPF	AGE
Number of outliers	1	0	3	0	3	3	4	0

Table 4.31

Splitting Point in Node 2

Variable	Num_preg	PGC	DBP	TRICEP	SERUM	BMI	DPF	AGE
Highest weighted average	0.6123	0.6326	0.6050	0.6110	0.6133	0.6371	0.6098	0.6309
Location of split	4th	19th	19th	25th	77th	28th SP: 26.2	232th	6th

Table 4.31 shows that the Gini purity index in node 2. Gini purity index are about the same in all variables. The highest Gini among all the variables are BMI with the value of 0.6371 where it is slightly higher than PGC (0.6326). The patients are split into node 4 and node 5 with the splitting point 26.2.

Table 4.32

Splitting Point in Node 3

Variable	Num_ preg	PGC	DBP	TRICEP	SERUM	BMI	DPF	AGE
Highest weighted average	0.6929	0.6907	0.6889	0.6908	0.6930	0.7189	0.7192	0.7362
Location of split	2th	12th	5th	27th	22th	6th	17th	34th SP: 59

The Gini purity index of node 3 is shown in Table 4.32. Node 3 contains lower complexity of node. Variable of AGE is selected with splitting point of 59 where the Gini index is 0.7362. Since the value of 0.7362 has achieved the threshold (Gini purity index of more than 0.7), therefore node 3 split into the final nodes (node 6 and node 7).

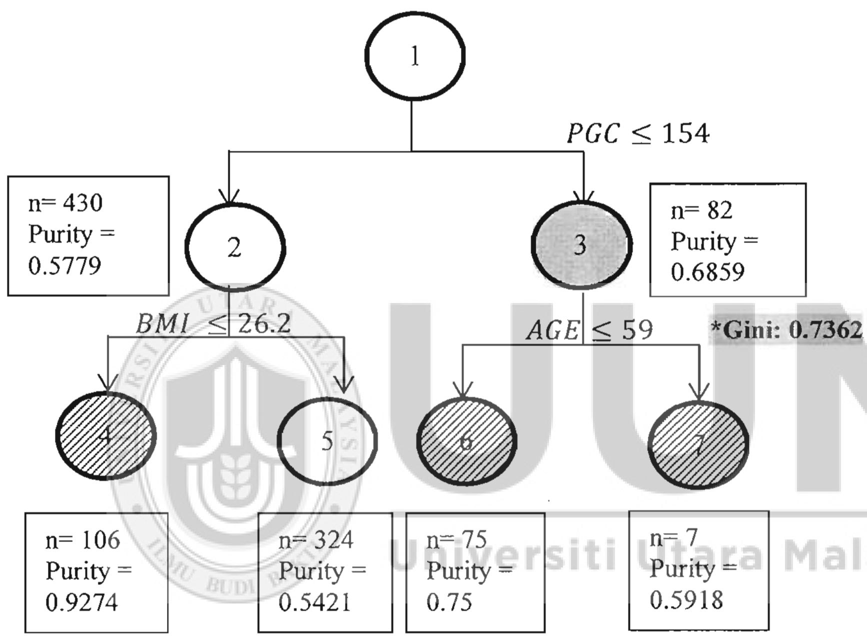


Figure 4.29. Child nodes from node 2 and node 3

Node 4 is terminal nodes since the overall purity of the node has already achieved the purity index (0.9274) in the node which is more than 0.7. Node 5 still impure (0.5421), more split need to been done to achieve the maximum homogeneity. All the processes are repeated until one of three thresholds are met.

The whole structure of traditional tree, pruned tree and Winsorize tree are displayed in Figure 4.30 to Figure 4.32.

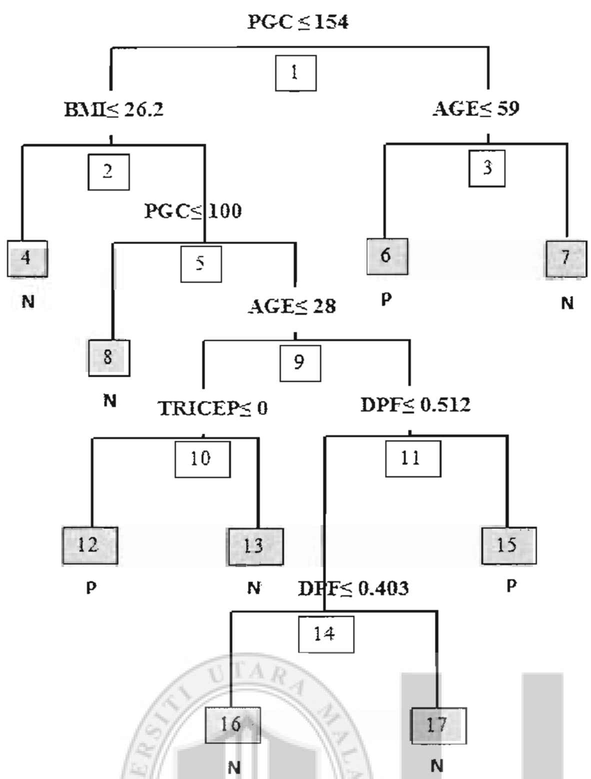


Figure 4.30. Winsorize tree of Pima Indians

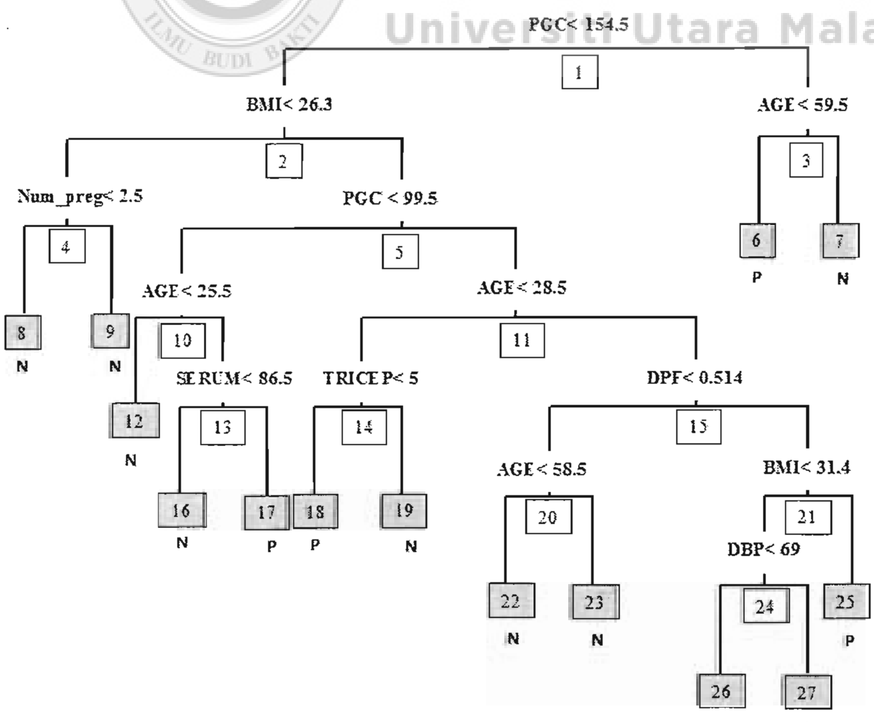


Figure 4.31. Traditional tree of Pima Indians

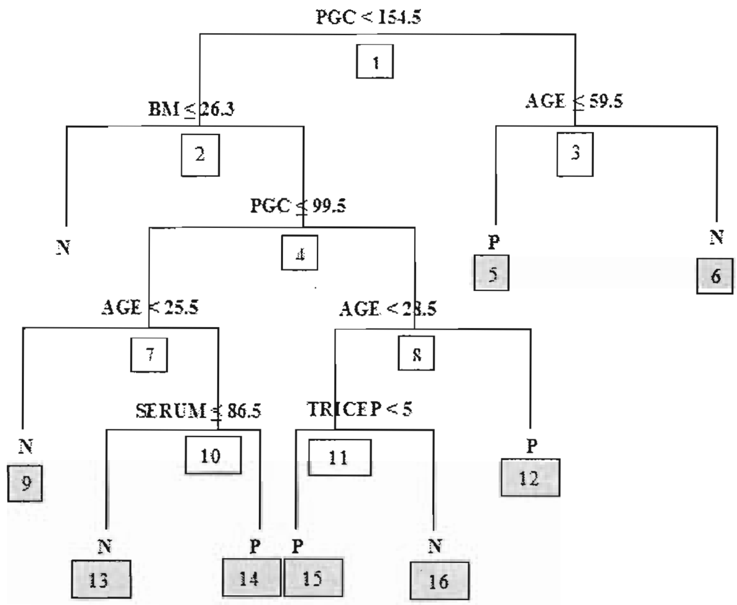


Figure 4.32. Pruned tree of Pima Indians

4.5.3 The Evaluation of Winsorize Tree for Pima Indians Data

After the completion of the trees, Comparison between trees is conducted to examine the performances of each tree.

Table 4.33

Comparison between Traditional Tree, Pruned Tree and Winsorize Tree

PIMA INDIANS:	**Traditional Tree**	**Pruned Tree**	**Winsorize Tree**
i. Number of splitting	13	8	8
ii. Number of leaves	14	9	9
iii. Number of variable use	8	5	5
iv. Name of variables used	1. PGC 2. BMI 3. AGE 4. Num_preg 5. SERUM 6. TRICEP 7. DPF 8. DBP	1. PGC 2. BMI 3. AGE 4. SERUM 5. TRICEP	1. PGC 2. BMI 3. AGE 4. SERUM 5. DPF

PIMA INDIANS:	Traditional Tree	Pruned Tree	Winsorize Tree
v. Error rate	**0.2188**	**0.2656**	***0.1758**
vi. Extreme value detected:			
a. First node	-		109
b. Second node	-		109
c. Third node	-		14
d. Fifth node	-		62
e. Ninth node	-		38
d. Tenth node	-		10
e. Eleventh node	-		22
f. Fourteenth node	-		15

Based on the result in Table 4.33, the number of splitting, the number of leaves and the variables used in pruned tree and Winsorize tree are similar which is far lower than the traditional tree. Although pruned tree and Winsorize tree are having same number of variable used, one of the variables used is different. For instance, pruned tree used SERUM but Winsorize tree used DPF as potential variables. Obviously, Winsorize tree perform better with lower error rate which is only 0.1758 compared to traditional tree and pruned tree.

Winsorize tree produced protection to all the potential outliers instead of removing or ignoring them. At least, the effect of outliers can be reduced to the minimum during the construction of tree. Moreover, time can be saved by not going through the process of pre-processing and pruning. In short, we have confident to say that Winsorize tree is more reliable in real life compared to the current tree even in big data set such as Pima Indians data set. At least, it is comparable to the traditional tree and pruned tree.

4.6 Case 4: Classification in Iris Data

Perhaps iris flower data set is one of the best and prominent case of study in pattern recognition literature. The Iris data was collected by Edgar Anderson in which the flowers were classified into 3 different species (Iris Setosa, Iris Virginica and Iris Versicolor). The data consists of 50 examples from each species and four variables were used in measurements which are the length and the width of Sepal and Petal. The data was popularised by Fisher in year 1936 as he developed the linear discriminant model to distinguish the species (Fisher, 1936; Duda & Hart, 1973). There are four variables of iris data set namely SepalLength (sepal length), SepalWidth (sepal width), PetalLength (petal length) and PetalWidth (petal width) to discriminate 3 classes which are Iris-setosa (Iris Setosa), Iris-versicolour (Iris Versicolour) and Iris-Virginica (Iris Virginica).

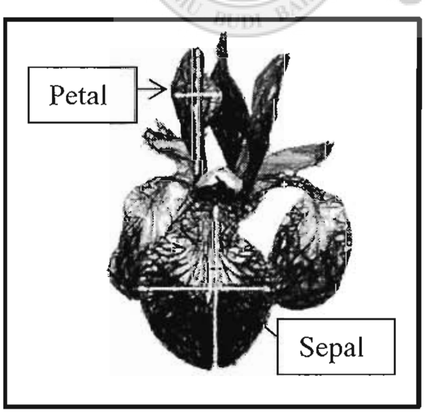


Figure 4.33. Iris flower

4.6.1 The Statistical Background of Iris Data

Table 4.34

Frequency Table of Iris Data Set

Class of Iris	**Iris-setosa**	**Iris-versicolour**	**Iris-Virginica**	**Total**
Frequency	30	33	37	100

Table 4.35

Statistical Description of Iris Data Set

Variables	**Mean**	**Median**	**Std. Deviation**	**Variance**	**Skewness**	**Kurtosis**
SepalLength	5.89	5.80	0.81	0.66	0.40	-0.50
SepalWidth	3.07	3.00	0.46	0.22	0.40	-0.18
PetalLength	3.82	4.40	1.73	3.01	-0.37	-1.28
PetalWidth	1.24	1.30	0.76	0.57	-0.17	-1.28

Iris data set consists of 150 samples of iris flowers. Two third of the observations are used for training set and the remaining are used for test set. In 100 training set selected randomly, 30 from the group of Iris-setosa, 33 from Iris-versicolour and the rest from Iris-Virginica. Table 4.36 summarises some estimated statistics from the training set, in which the mean and the median values are not much difference. Meanwhile, both skewness and kurtosis reflect that the distribution of each variable is somewhat symmetric hence the data may not badly be affected by the occurrence outliers.

To investigate the detail of the Iris data especially outliers, we plotted the distribution of class for each variable, distribution and separating point. If outlier is detected, then Winsorize method will be used to normalise the data.

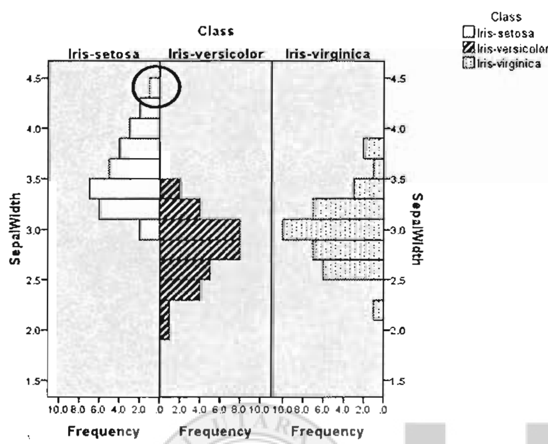


Figure 4.34(a). Original data of variable SepalLength

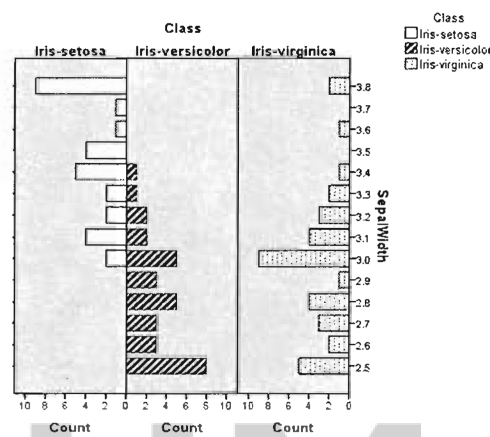


Figure 4.34(b). Winsorize data of variable SepalLength

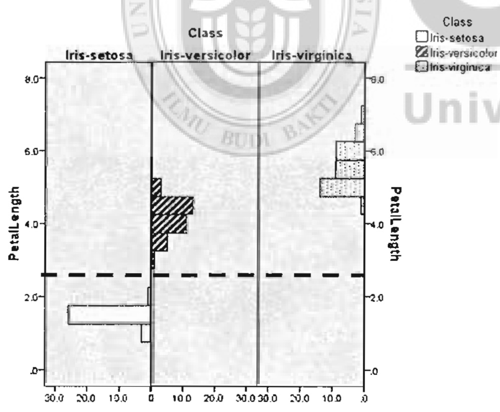


Figure 4.35. Original data of variable PetalLength

Since most of the classes are overlapping onto each other, it might be very hard to get a good splitting point. Thus, the Gini purity index is expected will be very low. Figure 4.35 shows a good cutting point (dotted line), where that cutting point can clearly separate three groups. However, the result is just based on our naked eye and personal judgment. In the following section, we will test again the outlier using boxplot and calculate the Gini purity index to get the best splitting criteria.

Table 4.36

Normality Tests

Variables	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SepalLength	0.086	100	**0.065**	0.973	100	**0.037**
SepalWidth	0.103	100	**0.010**	0.982	100	**0.183**
PetalLength	0.184	100	**0.000**	0.886	100	**0.000**
PetalWidth	0.157	100	**0.000**	0.912	100	**0.000**

To test whether the variable is normal distribution, Kolmogorov-Smirnov and Shapiro-Wilk test are carried out. We found that only the p-value in SepalLength and SepalWidth in Sharpiro-Wilk test are greater than 0.05, the rest are less than 0.05. Therefore, we can define that SepalLength and SepalWidth are probably normally distributed.

4.6.2 The Construction of Winsorize Tree for Iris Data

To ensure the occurrence of outliers, boxplot is plotted as in Figure 4.35. There are three outliers found in variable of SepalWidth which are objects 55, 24 and 91. Therefore, Winsorize need to be carried out before performing Gini purity index. Table 4.37 shows the total outliers found in each variable.

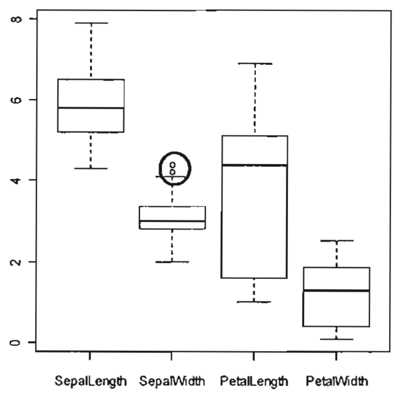


Figure 4.36. Outlier detection using boxplot

Table 4.37

Outliers in Parent Node

Variable	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
Number of outliers	0	3	0	0

10% of the data from SepalWidth has been Winsorize. Then, all the data must be sorted for measuring the Gini purity. The highest Gini purity index within and between the variable is chosen as potential variable for splitting with the cutting point. The process of calculation can referred to case I (Breast tissue). Table 4.38 shows Gini purity index between variable and their location of split in parent node.

Table 4.38

Splitting in Parent Node

Variable	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
Highest weighted average	0.5305	0.4619	0.6521	0.6511
Location of split	10th	9th	9th SP: 1.9	5th

Gini purity measurement shows that PetalLength scores the highest weighted average among all the variables with the index of 0.6521. The location of splitting is on the 9th with the splitting point of 1.9. Tree picture of parent node is displayed in Figure 4.37.

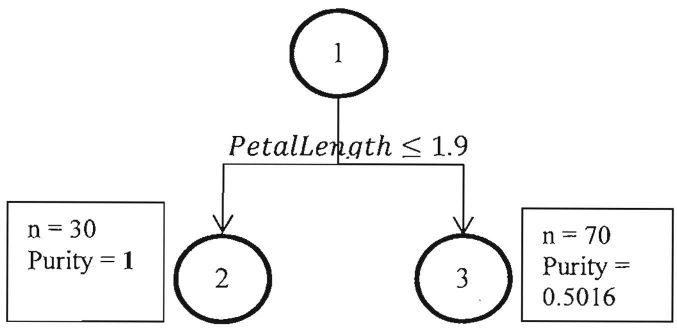


Figure 4.37. Child nodes from parent node

Table 4.39

Number of Observations in Node 2 and Node 3

Group / Node	Setosa	Versicolor	Virginica
Node 2	30	0	0
Node 3	0	33	37

In node 2, the overall Gini index is fully pure as only has one group in it. In contra, node 3 has 70 observations in it where 33 of Versicolor and 37 of Virginica. The overall Gini index in node 3 is 0.5016. Since node 3 has not reached the thresholds, further split is needed using the original data set.

In node 3, no outlier has been detected which means that all the data are under the acceptable range. Gini purity index is performed again for the next split.

Table 4.40

Splitting Point in Node 3

Variable	**SepalLength**	**SepalWidth**	**PetalLength**	**PetalWidth**
Highest weighted average	0.6331	0.5356	0.8983	0.8983
Location of split	12th	9th	14th SP: 4.8	7th

The next potential of splitting points is either PetalLength or PetalWidth with Gini purity index of 0.8983. Practitioner can choose any one of them if the Gini purity index is equal. In this case we choose PetalLength with the splitting 4.8. Due to the Gini purity index has achieved the threshold (Gini purity index within variable is more than 0.7), we considered node 3 is having the last split with node 4 and node 5 as terminal nodes. This is the splitting rule that we introduced in this study. Tree structure is displayed in Figure 4.38.

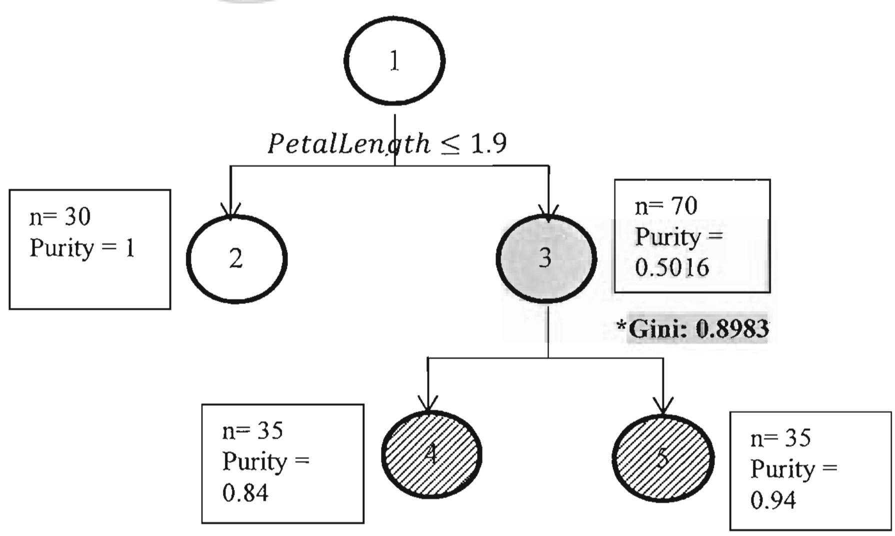


Figure 4.38. Child nodes from node 3

Table 4.41

Number of Observations in Node 4 and Node 5

Group / Node	Setosa	Versicolor	Virginica
Node 4	0	32	3
Node 5	0	1	34

In Table 4.41, there are 35 observations in both node 4 and node 5 respectively. Node 4 contains 0 Setosa, 32 Versicolor and 3 Virginica while node 5 contains 0 Setosa, only 1 Versicolor and 34 Virginica.

Winsorize tree produced a tree which resembles the traditional tree. The different is the first splitting point as traditional tree used the midpoint of each consequence point. But, it does not affect much on the result. Due to the size of tree, pruning tree is not allowed. Therefore the pruned tree will be the same as the original tree. The structures of trees are depicted in Figure 4.39, Figure 4.40 and Figure 4.41.

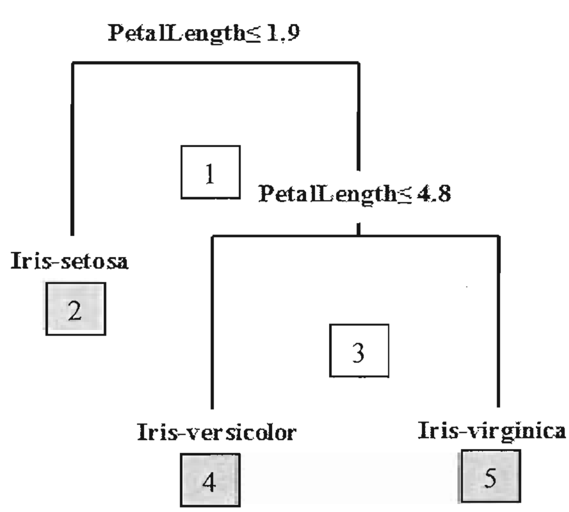


Figure 4.39. Winsorize tree of Iris

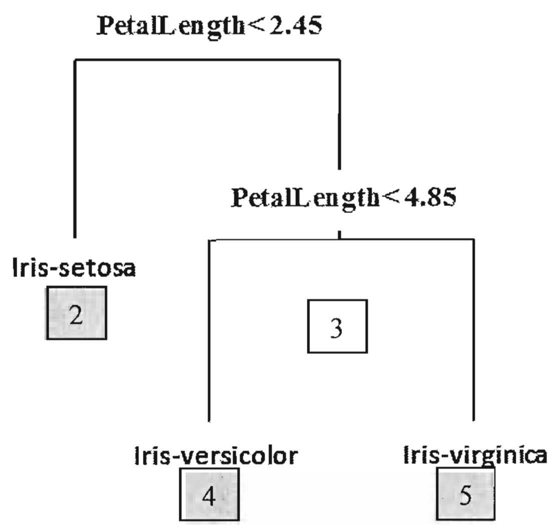


Figure 4.40. Traditional tree of Iris

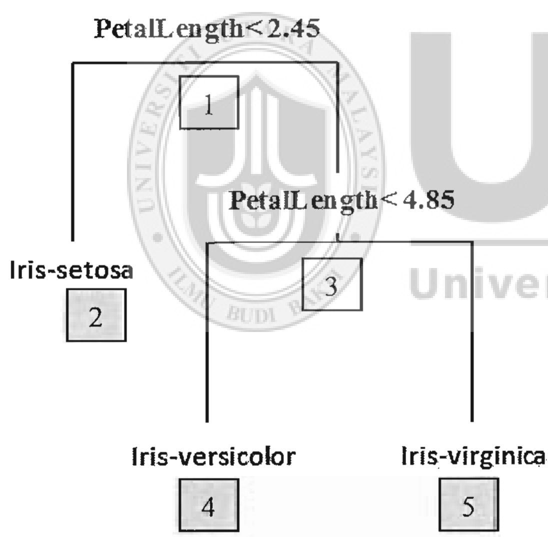


Figure 4.41. Pruned tree of Iris

4.6.3 The Evaluation of Winsorize Tree for Iris Data

The structure showed that there are no different between the trees. Only the cutting point in Winsorize tree is slightly different from the traditional tree. The details of comparison between trees are shown in Table 4.42.

Table 4.42

Comparison between Traditional Tree, Pruned Tree and Winsorize Tree

IRIS:	**Traditional Tree**	**Pruned Tree**	**Winsorize Tree**
i. Number of splitting	2	2	2
ii. Number of leaves	3	3	3
iii. Number of variable use	1	1	1
iv. Name of variable used	1. PetalLength	1. PetalLength	1. PetalLength
v. Error rate	**0.06**	**0.06**	**0.06**
vi. Extreme value detected			
a. First node	-	-	3
b. Second node	-	-	0

In fact, there is no different between the trees. We purposely try only this data set to show that when the data consist one or few outliers or even no outlier at all, the proposed Winsorize tree is still stay stable which the result is comparative to the traditional tree and pruned tree. Moreover, at least some potential outliers have been found and penalised using Winsorize tree instead of ignoring them. The same score on error rate perhaps is best explained by the fact that only a variable was used to discriminate the classes. It means, regardless of different type of trees (e.g. CHAID, ID3 etc.) use for such like this case, the error rate will be the same, and this error rate is the lowest.

4.7 Case 5: Classification in Bumpus Sparrow Data

In year 1898, a severe winter storm near Providence happened causing some of the local sparrow died and some survive. Herman Bumpus decided to investigate on theory of evolution based on some physical characteristics such as total length of humerus, length of bead and head, alar length, total length, and keel of sternum. The response variable is either survives or died. Total observations collected are 49 where 33 observations are used as training set. We purposely try on this small data set to see that whether the new model is comparative to the traditional one. The independent variables used in this data are `Length_humerus` (length of humerus), `Length_bead_head` (Length of bead and head), `Alar_length` (Alar length), `Total_length` (Total length) and `Length_keel_sternum` (length of keel and sternum) and the response variable are either `S` (survive) or `D` (died).

4.7.1 The Statistical Background of Bumpus Sparrow Data

Table 4.43

Frequency Table of Bumpus Sparrow Data Set

Class of Iris	S	D	Total
Frequency	13	20	33

Bumpus sparrow data has a very small number of observations which the training set is only 33 where 13 from the group of survive and 20 from the group of died.

Table 4.44

Statistical Description of Bumpus Sparrow Data Set

Variables	Mean	Median	Std. Deviation	Variance	Skewness	Kurtosis
Total_length	158.30	158.00	3.60	12.97	0.05	-0.99
Alar_length	241.76	242.00	5.41	26.45	0.11	-1.03
Length_bead_head	31.56	31.50	0.79	0.63	0.46	-0.45
Length_humerus	18.57	18.60	0.59	0.35	-0.06	-0.20
Length_keel_sternum	20.88	20.80	0.96	0.93	0.30	-0.62

According to Table 4.44, the skewness of all the variables is in the range of -1 and 1 which mean that the curve is symmetry (approximately 0 skewness). Besides, we also gain negative values in kurtosis for all the variables. The value gains are just slightly shifted from 0. It means that the distribution is flatter. Alar length produces the highest variance where the value is 26.45. To ensure whether the data contains outlier or not, we plotted distribution graph. Through the graph, we can also spot the separation of the class of "S" and "D".

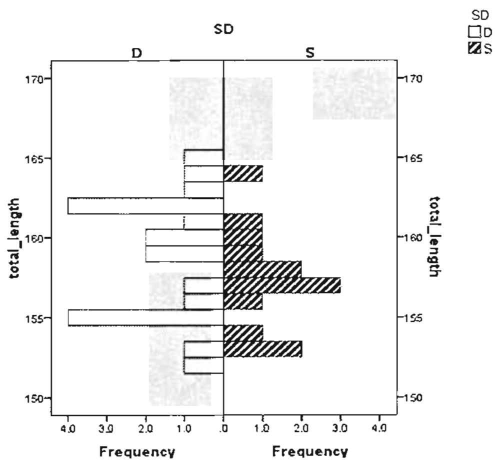


Figure 4.42(a). Original data of variable Total_length

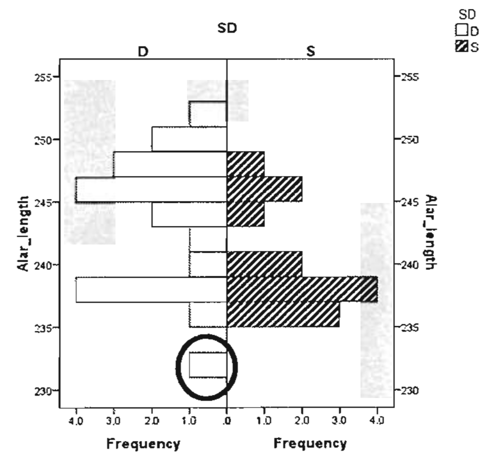


Figure 4.42(b). Original data of variable Alar_length

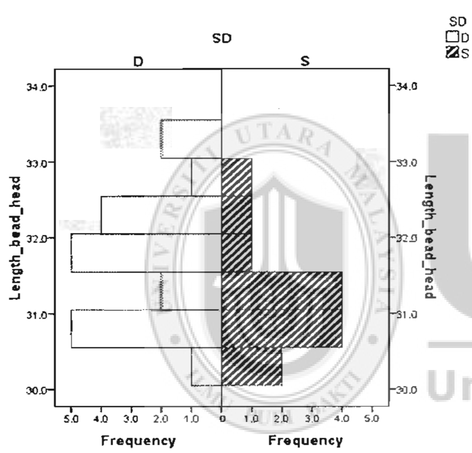


Figure 4.42(c). Original data of variable Length_bead_head

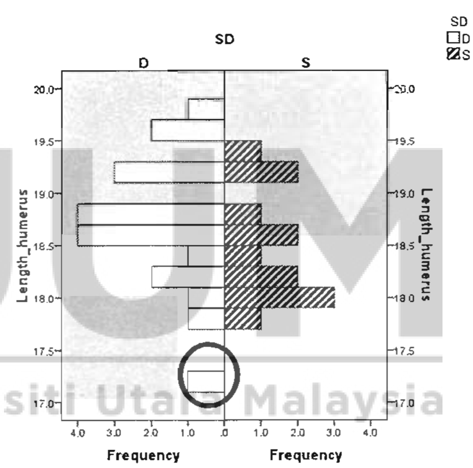


Figure 4.42(d). Original data of Length_humerus

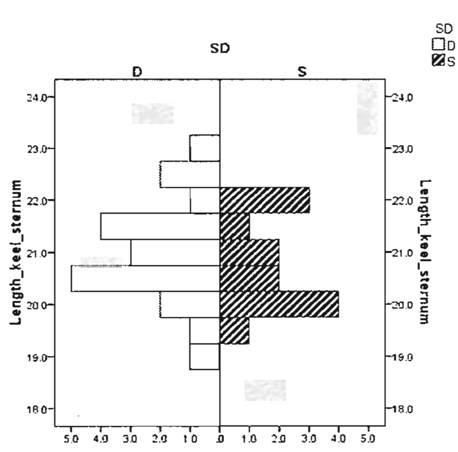


Figure 4.42(e). Original data of variable Length_keel_sternum

Based on the graph, Figure 4.42(a), Figure 4.42(c) and, Figure 4.42(e) clearly show that the data has no outliers. However, Figure 4.42(b) and Figure 4.42(d) show that there are some potential outliers located on the floor of the data. However, further analysis need to be done to ensure whether the seen outliers are true.

Table 4.45

Normality Tests

Variables	**Kolmogorov-Smirnov**[a]			**Shapiro-Wilk**		
	Statistic	df	Sig.	Statistic	df	Sig.
Total_length	0.096	33	**0.200**	0.966	33	**0.381**
Alar_length	0.161	33	**0.029**	0.957	33	**0.215**
Length_bead_head	0.121	33	**0.200**	0.966	33	**0.380**
Length_humerus	0.089	33	**0.200**	0.986	33	**0.936**
Length_keel_sternum	0.116	33	**0.200**	0.973	33	**0.579**

Table 4.45 shows the normality test. All the variables in both test show that the p-values are more than 0.05. Therefore, we can conclude that the data is normal.

4.7.2 The Construction of Winsorize Tree for Bumpus Sparrow Data

Due to some suspicious value displayed in Figure 4.42(b) and Figure 4.42(d), boxplot has been conducted to investigate the existing outlier.

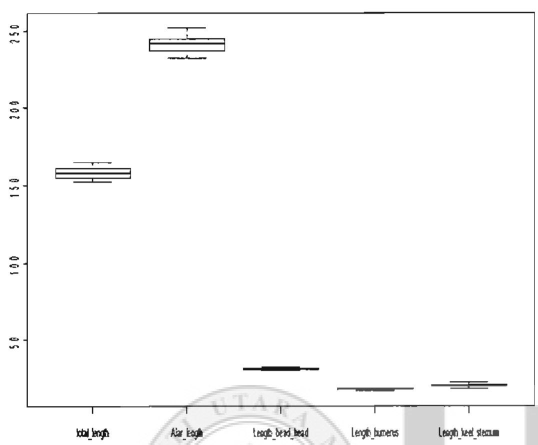


Figure 4.43. Outlier detection using boxplot in parent node

Table 4.46

Outliers in Parent Node

Variable	Total_length	Alar_length	Length_bead_head	Length_humerus	Length_keel_sternum
Number of outliers	0	0	0	0	0

According to the boxplot measurement as in Figure 4.43 and the result in Table 4.46, no outlier is detected. It means that all the values are in the bound. Therefore, the Winsorize process can be skipped in this stage. Next, Gini purity measurement is performed after sorting all the values.

Table 4.47

Splitting Point in Parent Node

Variable	Total_length	Alar_length	Length_bead_head	Length_humerus	Length_keel_sternum
Highest weighted average	0.5688	0.5759	0.5852	0.5535	0.5653
Location of split	10th	6th	11th SP: 31.5	12th	18th

Based on Table 4.47, all the variables are comparative where the weighted average is about 0.55. However, the variable of Length_bead_head gains the highest weighted average among all the variables with the index of 0.5852 with the splitting point of 31.5. Tree picture of parent node is displayed in Figure 4.44.

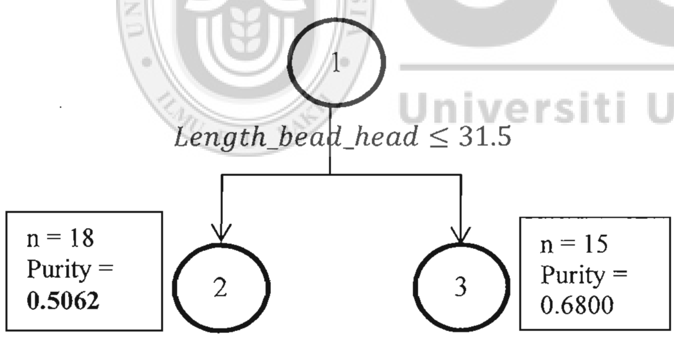


Figure 4.44 Child nodes from parent node

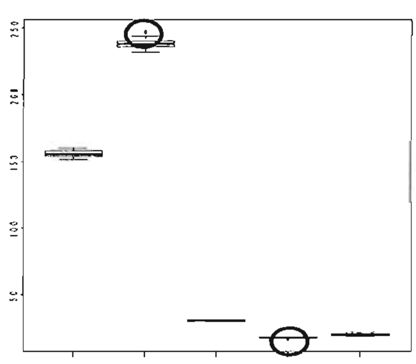
Table 4.48

Number of Observations in Node 2 and Node 3

Group / Node	S	D
Node 2	10	8
Node 3	3	12

In Table 4.48, Node 2 shows that the group of S is slightly higher than the group of D whereas in node 3, the number of group D is 4 times higher than the group of S. Since the purity in both nodes is still not approaching the thresholds, further split is necessary.

In node 2 and node 3, the original data is again to be investigated for the existence of outlier.



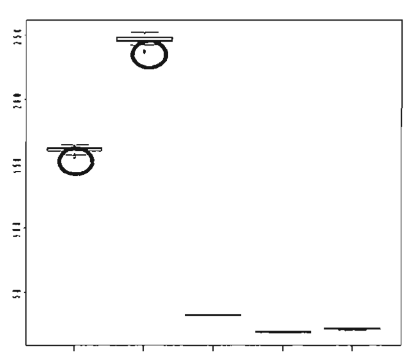


Figure 4.45. Outlier detection using boxplot in node 2(left) and node 3(right)

From the boxplot (in Figure 4.45), we can only see roughly about the presence of outliers. Based on the calculation in both nodes, node 2 has 2 outliers while node 3 has 4 outliers as shown in Table 4.49 and Table 4.50 respectively. Therefore, Winsorize method needs to be performed to neutralise the heavy tail before measuring the Gini purity index. The result of Gini purity index is shown in Table 4.51 and Table 4.52.

Table 4.49

Outlier in Node 2

Variable	**Total_length**	**Alar_length**	**Length_bead _head**	**Length_ humerus**	**Length_keel _sternum**
Number of Outliers	0	1	0	1	0

Table 4.50

Outlier in Node 3

Variable	**Total_length**	**Alar_length**	**Length_bead _head**	**Length_hu merus**	**Length_keel _sternum**
Number of outliers	1	1	0	2	0

Table 4.51

Splitting Point in Node 2

Variable	Total_length	Alar_length	Length_bead _head	Length_ humerus	Length_keel_ sternum
Highest weighted average	0.6049	0.5259	0.5425	0.5278	0.5852
Location of split	4th SP: 155	1[st]	1st	7th	11th

Based on the highest weighted average in Table 4.51, the potential variable to be chosen in node 2 is Total_length with the splitting point 155. Therefore, the objects are split to node 4 and node 5. There are 9 objects assigned to node 4 where 3 are survival and 6 are dead. 9 objects are assigned into node 5 where 7 from the group of survival and 2 from the group of dead. The total purity in node 4 and node 5 are 0.5556 and 0.6543 respectively. Since the thresholds are still unachievable, further split is needed.

Table 4.52

Splitting Point in Node 3

Variable	Total_length	Alar_length	Length_bead _head	Length_hu merus	Length_keel _sternum
Highest weighted average	0.6929	0.7000	0.7200	0.7333	0.7714
Location of split	3rd	2nd	3rd	2nd	1st SP: 20

In Table 4.52, we can spot that all the variables are about 0.7. The highest weighted average is Length_keel_sternum which is 0.7714. Since the value has already achieved the threshold(> 0.7), the last split is allowed from node 3 to split into node 6 and node 7 before stop splitting. There are only one object in node 6 which is survival and 14 objects in node 7 (12 from the group of dead and 2 from the group of survival). The splitting point is 20. Second level of split is displayed in Figure 4.46.

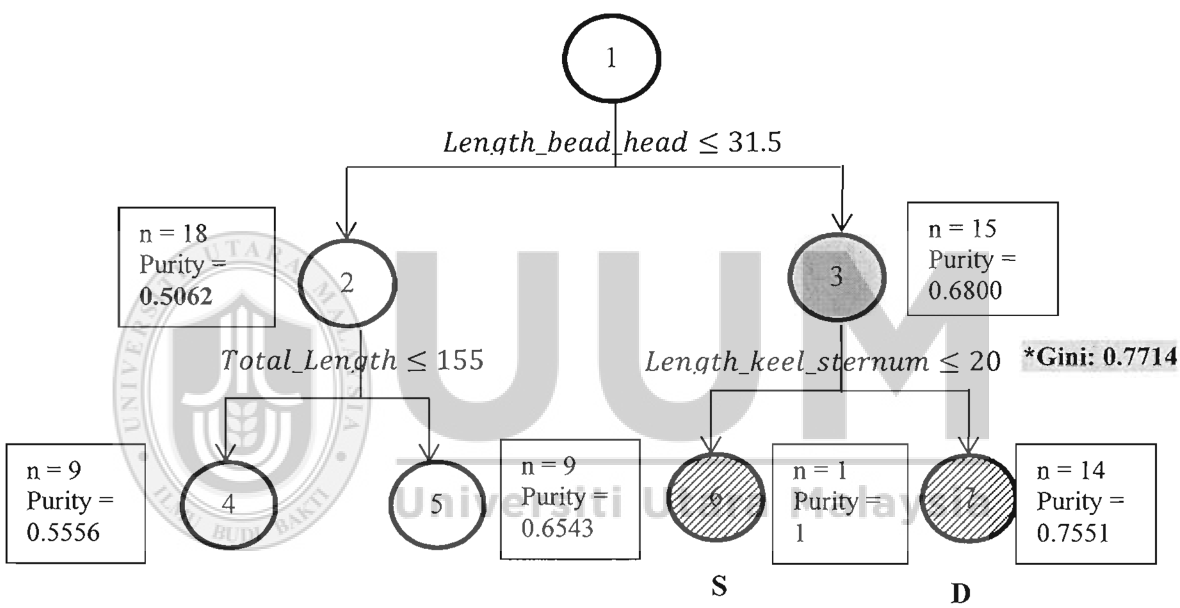


Figure 4.46. Child Nodes from Parent Node 2 and Node 3

Since node 4 and node 5 haven't reached the threshold so further split need to be run. In node 4, no outlier is detected. However, in node 5 there is one outlier found in variable Alar_length. Therefore, Winsorize method has to carry out to neutralise the heavy tail. Table 4.54 shows the number of outlier detected in node 5.

Table 4.53

Number of Observations in Node 4, Node 5, Node 6 and Node 7

Group / Node	D	S
Node 4	6	3
Node 5	2	7
Node 6	0	1
Node 7	12	2

Table 4.54

Outlier in Node 5

Variable	Total_length	Alar_length	Length_bead_head	Length_humerus	Length_keel_sternum
Number of outliers	0	1	0	0	0

Table 4.55

Splitting in Node 4

Variable	Total_length	Alar_length	Length_bead_head	Length_humerus	Length_keel_sternum
Highest weighted average	0.7333	0.6190	0.6667	0.6667	0.5833
Location of split	3rd SP: 153	2nd	3rd	5th	7th

Table 4.55 shows the Gini purity index in node 4. The variable of Total_length gains the highest weighted average with the value of 0.7333 and the splitting point is 153. According to the threshold, due to the Gini purity index has reached above 0.7; node 4

is only allowed to split for the final nodes, which are node 8 and node 9. Node 8 contains 4 objects (2 from survival and 2 from dead) whereas node 9 contains 5 objects which all only 1 from the group of survival and the rest are the group of dead.

Table 4.56

Splitting Point in Node 5

Variable	Total_length	Alar_length	Length_bead_head	Length_humerus	Length_keel_sternum
Highest weighted average	0.7778	0.8058	0.6667	0.7037	0.8056
Location of split	3rd	6th SP: 238	1st	3rd	1st

According to the result in Table 4.56, the highest weighted average is Alar_length with Gini purity index 0.8058. The splitting point is 238. Since the value has achieved the threshold (above 0.7); the next split from node 5 will be the terminal nodes (node 10 and node 11). There are 2 objects from the group of survival and 1 object from the group of dead in node 10. Therefore, node 10 is classified as group D. In contra, 6 objects are assigned in node 11 which all are the group of dead. Full Winsorize tree can be referred to Figure 4.47.

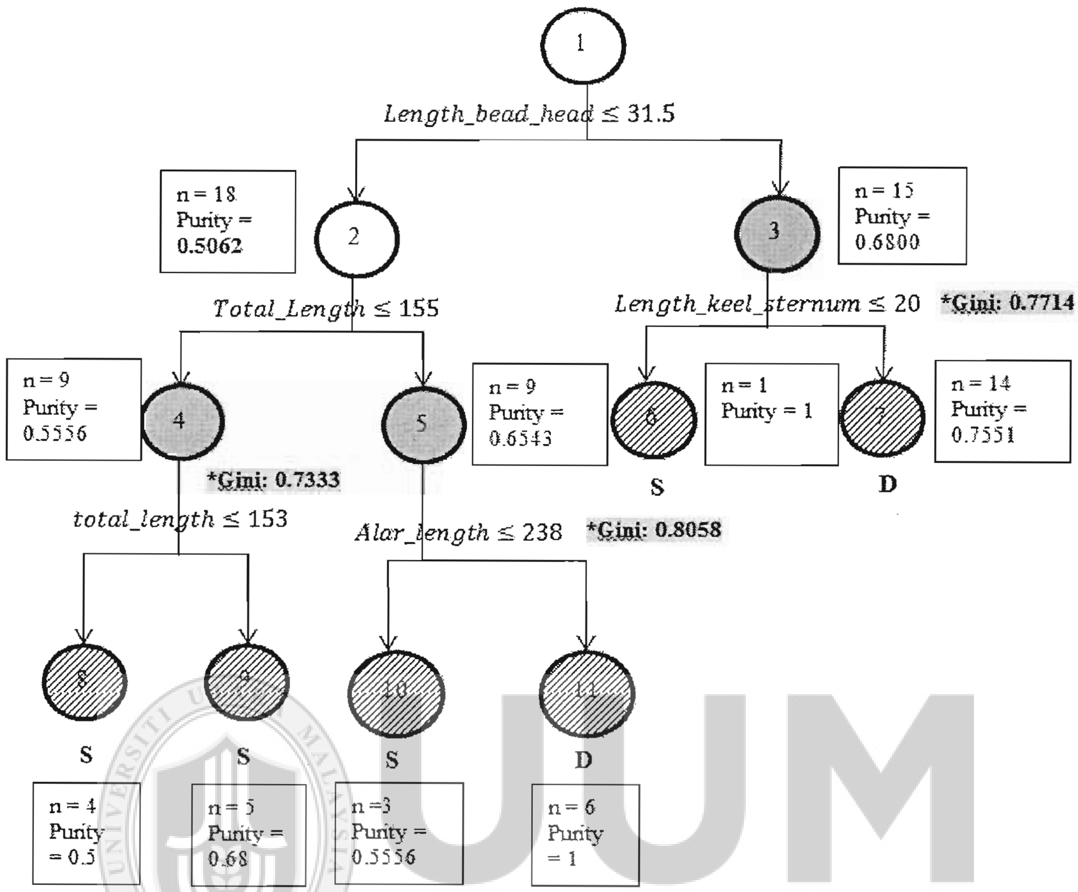


Figure 4.47. Child nodes from node 4 and node 5

Table 4.57

Number of Observations in Node 8, Node 9, Node 10 and Node 11

Group / Node	D	S
Node 8	2	2
Node 9	4	1
Node 10	1	2
Node 11	6	0

Comparison between three types of tree is shown in Figure 4.48 to Figure 4.50.

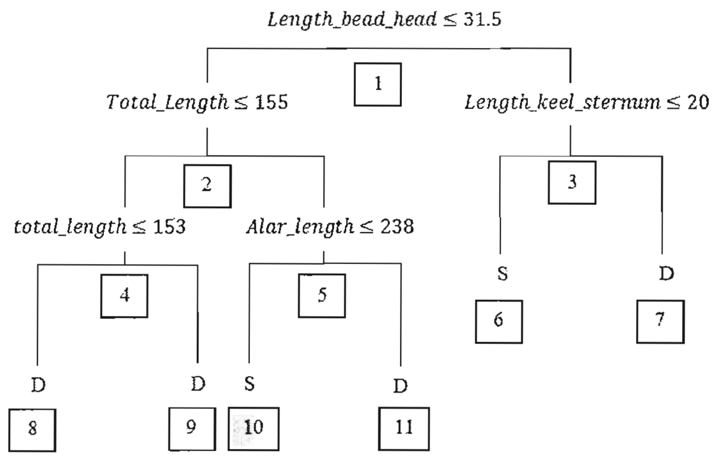


Figure 4.48. Winsorize tree of Bumpus Sparrow

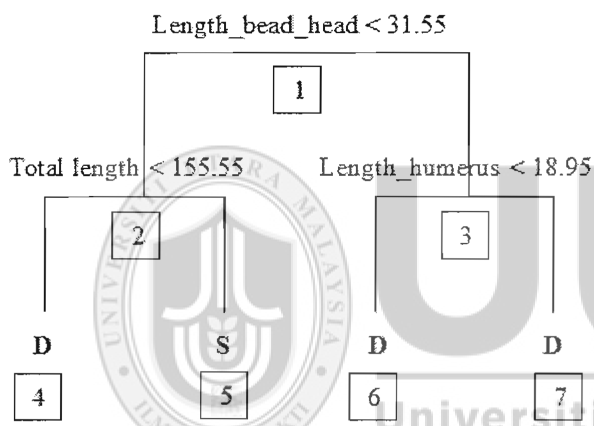


Figure 4.49. Traditional tree of Bumpus Sparrow

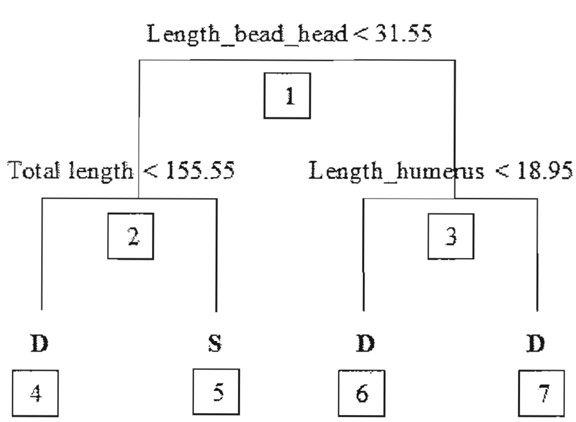


Figure 4.50. Pruned tree of Bumpus Sparrow

4.7.3 The Evaluation of Winsorize Tree for Bumpus Sparrow Data

Table 4.58 shows the comparison of the performance between traditional tree, pruned tree and Winsorize tree.

Table 4.58

Comparison between Traditional Tree, Pruned Tree and Winsorize Tree

BUMPUS SPARROW:	**Traditional Tree**	**Pruned Tree**	**Winsorize Tree**
i. Number of splitting	3	3	5
ii. Number of leaves	4	4	6
iii. Number of variable used	3	3	4
iv. Name of variables used	1. Length_bead_head 2. Total length 3. Length_humerus	1. Length_bead_head 2. Total length 3. Length_humerus	1. Length_bead_head 2. Total length 3. Length_keel_sternum 4. Alar_length
4. Error rate	**0.6875**	**0.6875**	***0.5625**
5. Outliers detected:			
a. First node	-	-	0
b. Second node	-	-	2
c. Third node	-	-	4
d. Fourth node	-	-	0
e. Fifth node	-	-	1

Based on the result we gain from the analysis, once again Winsorize tree surmount between the trees. Traditional tree and pruned tree contain the same result as the traditional tree is too small to be pruned. Both are having 3 number of splitting with 4 leaves. The numbers of variables used are only 3. In contrary, Winsorize tree contains one level more than traditional tree with 5 numbers of splitting and 6 numbers of leaves. Beside, Winsorize tree used four variables to construct the tree and it gains the lowest error rate which is 0.5625 compared to the others. According to our observation, we found that the high error rate in traditional tree is due to the masking variable. For instance, Alar_length is not used in traditional tree but this variable is vital to separate the class of objects as in Winsorize tree. In node 5 (Figure 4.46), this variable can successfully separate the objects into pure node as in node 11 (all are group D). In addition, the effect of outliers all are neutralized which produces a more precise and accurate tree for classification.

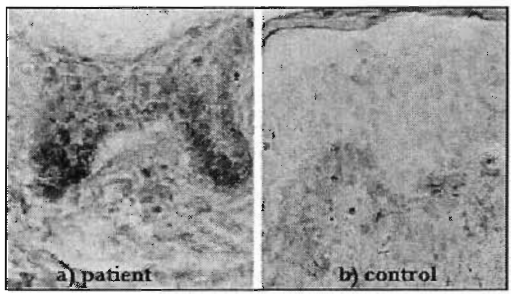
4.8 Case 6: Classification in Indians Liver Patient Dataset (ILPD)

Nowadays, patients with liver disease are increasing due to drugs, contaminated foods, alcohol consumption, inhalation of harmful gases and so forth. Therefore many classification techniques have been widely used in medical field to diagnose this problem (Jayakrisharan, Rajan, Jagdish & Sanjay, 2014).

Indians Liver Patient Dataset (ILPD) was collected from north east of Andhra Pradesh, India. Many research conducted used this data for comparative analysis and trying to improve in prediction accuracy (Ramana, Babu & Venkateswarlu, 2012). The data of ILPD is taken from UCI repository where the data contains 583 observations. The data has 10 independent variables and a dependent variable with two groups. There are 441 male patients and 142 females in record. 416 of the patients have liver problem and 167 have no liver patients in the group. Below are the data descriptions.

1. `Age` (Age of the patient)
2. `Gender` (Gender of the patient)
3. `TB` (Total Bilirubin)
4. `DB` (Direct Bilirubin)
5. `Alkphos` (Alkaline phosphotase)
6. `Sgpt` (Alamine Aminotransferase)
7. `Sgot` (Aspartate Aminotransferase)
8. `TP` (Total proteins)
9. `ALB` (Albumin)
10. `A/G` (Albumin and Globulin ratio)
11. `Class`
    i. `LP` (liver patient)
    ii. `NLP` (non liver patient)

From the data, 390 training set are selected randomly from the data and the remaining 193 data are selected as test set. As previous cases, preamble analysis has been carried out for better understanding about the data. Figure 4.51 shows the picture of Indian Liver Patient.



(a) (b)

Figure 4.51. Indians Liver Picture where a) patient and b) control

4.8.1 The Statistical background of ILPD

Table 4.59 and Table 4.60 show the statistical background of the data

Table 4.59

Frequency Table of Indians Liver Patient Dataset

Class	LP	NLP	Total
Number of patients	416	167	583

Table 4.60

Statistical Description of Indians Liver Patient Dataset

Variable	Mean	Median	Std. Deviation	Variance	Skewness	Kurtosis
Age	44.54	45.00	16.53	273.20	-0.04	-0.64
TB	2.80	1.00	5.58	31.18	7.13	75.83
DB	1.20	0.30	2.18	4.76	3.39	12.46
Alkphos	295.44	209.00	246.19	60611.05	3.85	18.88
Sgpt	87.07	35.00	212.22	45037.46	5.98	40.04
Sgot	115.64	39.00	340.42	115882.77	9.54	116.58

Variable	Mean	Median	Std. Deviation	Variance	Skewness	Kurtosis
TP	6.46	6.50	1.11	1.23	-0.21	0.19
ALB	3.20	3.20	0.80	0.65	-0.03	-0.45
AG	0.97	1.00	0.29	0.08	0.41	0.22

According to Table 4.60, there are big spread of the standard deviation in the data except ALB, TP and AG. The skewness of variable TB, Sgpt and Sgot are considered high skewed to the right as the positive value is quite high. Leptokurtic happened in TB, DB, Alkphos, Sgpt and Sgot as the value is exceeded 3. Thus, the information indicated about the existence of outliers in most of the variables.

We also plot the distribution of each class to spot the behaviour of the class as in Figure 4.52, Figure 4.53 and Figure 4.54 so that we get an idea on the potential cutting point which produces a good separation between the classes of the patients.

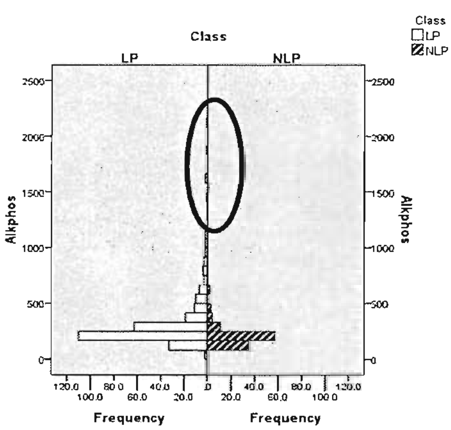


Figure 4.52(a). Original data of variable Alkphos

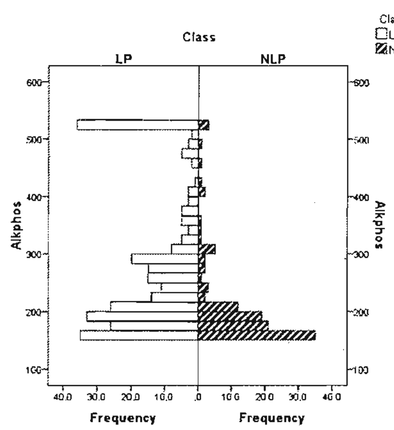


Figure 4.52(b). Winsorize data of variable Alkphos

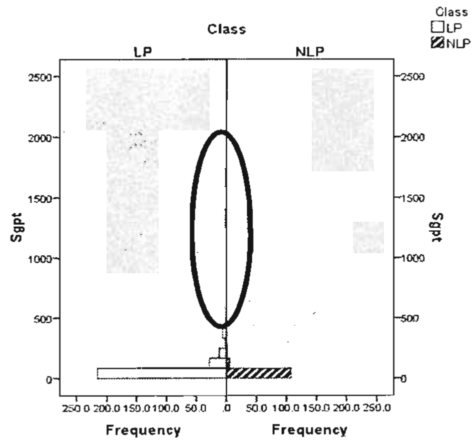


Figure 4.53(a). Original data of variable Sgpt

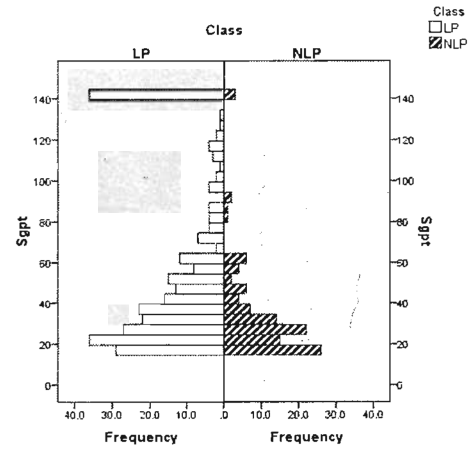


Figure 4.53(b). Winsorize data of variable Sgpt

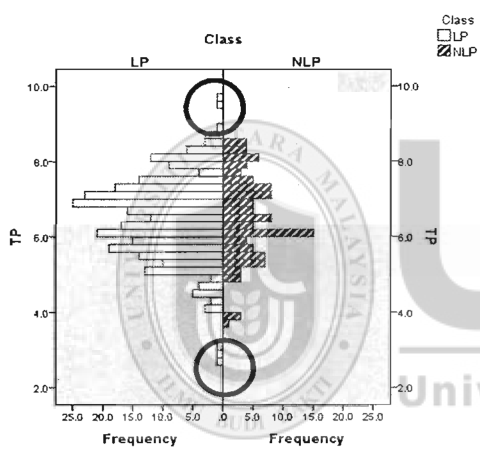


Figure 4.54(a). Original data of variable TP

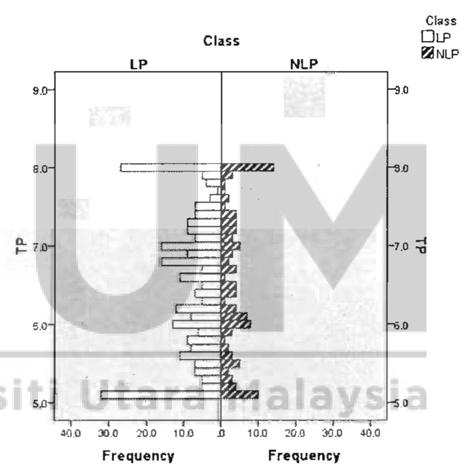


Figure 4.54(b). Winsorize data of variable TP

Three original data distribution histograms are plotted in Figure 4.52(a), Figure 4.53(a) and Figure 4.54(a). Based on the distribution in Figure 4.52(a) and Figure 4.53(a), we can identify that the data are having a long tails and there are probably consist outliers in the data. Winsorize method is applied to penalise those heavy tails so that the value are dragged to the acceptable range. Another problem can be seen in the Figures are the redundancies of group making it hard to be separated clearly.

Table 4.61

Normality Tests

	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	0.055	390	**0.007**	0.989	390	**0.006**
TB	0.309	390	**0.000**	0.519	390	**0.000**
DB	0.307	390	**0.000**	0.541	390	**0.000**
Alkphos	0.260	390	**0.000**	0.597	390	**0.000**
Sgpt	0.354	390	**0.000**	0.329	390	**0.000**
Sgot	0.372	390	**0.000**	0.269	390	**0.000**
TP	0.049	390	**0.023**	0.993	390	**0.060**
ALB	0.063	390	**0.001**	0.991	390	**0.016**
AG	0.125	390	**0.000**	0.948	390	**0.000**

According to normality test, the result in both test shows that all the variables are not normally distributed as all the p-value are less than 0.05 except `TP` in Sharpiro-Wilk test which 0.06 is slightly higher than 0.05.

4.8.2 The Construction of Winsorize Tree for ILPD

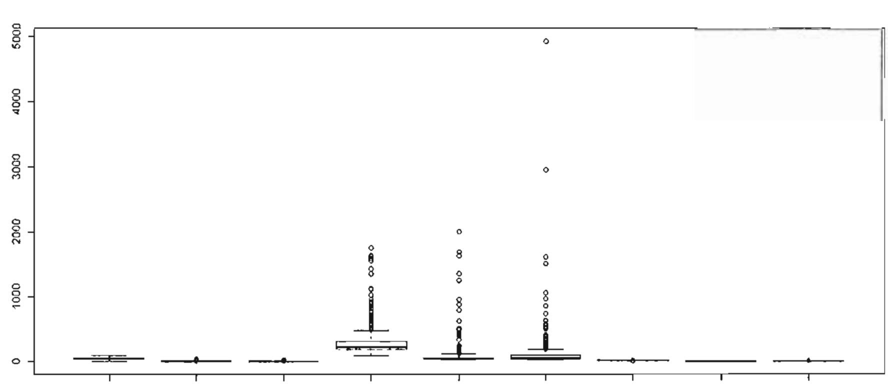


Figure 4.55. Outlier detection using boxplot

According to Figure 4.55, 267 outliers have been detected in this data where DB contains the highest number of outliers which is 57 outliers in it. Sgot has the longest tail where the highest value is 4929. It means that this value is shifted extreme far from the mean, 120.21. Such result explains why Table 4.61 showed that the standard deviation of Sgot is the highest (340.42) among all the variables. In contra, no outliers are detected in variable Age and ALB. Details on the number of outliers for each variable are recorded in Table 4.63.

Table 4.62

Outliers in Parent Node

Variable	Age	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	AG
Number of Outliers	0	56	57	47	50	45	5	0	7

The process of handling outliers using Winsorize method is similar to the one discussed in subsection 4.1.2 where those identified outliers have to be Winsorize prior to the computation of the Gini purity measurement.

Table 4.63

Splitting Point in Parent Node

Variable	Age	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	AG
Highest weighted average	0.6072	0.6331	0.6330	0.6274	0.6254	0.6305	0.5967	0.6132	0.6066
Location of split	22	10th SP: 1.3	10	52	15	44	16	23	21

The result displayed in Table 4.63 shows that variable TB has the highest Gini purity index (0.6331) where the splitting location is located at the tenth. The splitting point of TB is 1.3. Objects less than or equal to 1.3 is assigned to the left node whereas the rest is assigned to the right node. Left node contains 227 objects where 136 are from the group of LP and 91 from the group of NLP while right node contains 163 objects where about 88% of the objects come from the group of LP. The structure of the first split and the total number of patients are shown in Figure 4.56 and Table 4.65.

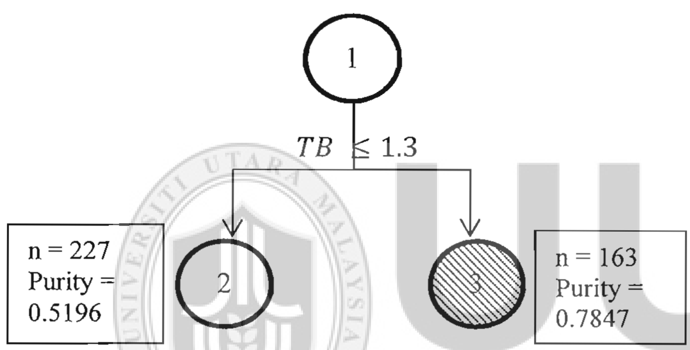


Figure 4.56. Child nodes from parent node

Table 4.64

Number of Patients in Node 2 and Node 3

Node \ Group	LP	NLP
Node 2	136	91
Node 3	143	20

The overall Gini purity index of node 2 and node 3 are 0.5196 and 0.7847. Due to the overall Gini purity index in node 3 is exceeding 0.7 (threshold), node 3 is considered as terminal node.

In Node 2, the process above is repeated in second level using the original data to get for the next binary nodes (node 4 and node 5). There are 170 outliers are found. However, no outlier is detected in variable `Age` and `ALB`. In contra, DB contains the highest number of outliers which is 94 outliers. The details are shown in Table 4.65.

Table 4.65

Number of Outliers in Node 2

Variable	Age	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	AG
Number of Outliers	0	7	94	16	21	25	3	0	4

Table 4.66

Splitting Point in Node 2

Variable	Age	TB	DB	Alkphos	Sgpt	Sgot	TP	ALB	AG
Highest weighted average	0.5389	0.5259	0.5210	0.5409	0.5428	0.5388	0.5272	0.5326	0.5281
Location of split	15	5	1	42	17th SP: 26	38	1	19	16

According to the Gini purity index in node 2, the highest one is `Sgpt` which as shown in Table 4.66. With the splitting of 26, node 2 split the data into node 4 and node 5. There are 105 objects assigned to node 4 where 57 are come from the group of LP and 48 from the group of NLP. The overall purity index in that node is 0.5037. Since no threshold has been met, future splitting is needed. Besides, node 5 contains 122 objects where 79 and 43 from the group of LP and NLP respectively. The overall purity index in it is 0.5435. Again, the node needs to be process for the next level due to unachievable of any of the thresholds.

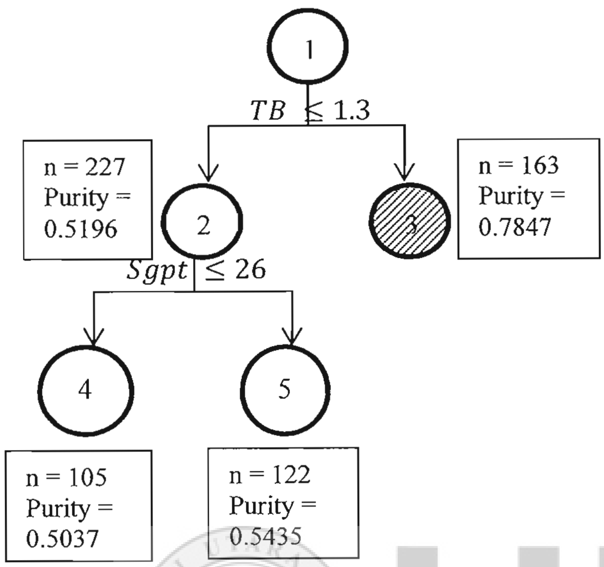


Figure 4.57. Child nodes from node 2

Table 4.67

Number of Observations in Node 3, Node 4 and Node 5

Node \ Group	LP	NLP
Node 3	143	20
Node 4	57	48
Node 5	79	43

The process is repeated recursively until one of the threshold is reached. The final structure of Winsorize tree is shown in Figure 4.58. And, traditional tree and pruned tree are shown in Figure 4.59 and Figure 4.60. The overall assessments of all trees are discussed in the Table 4.68.

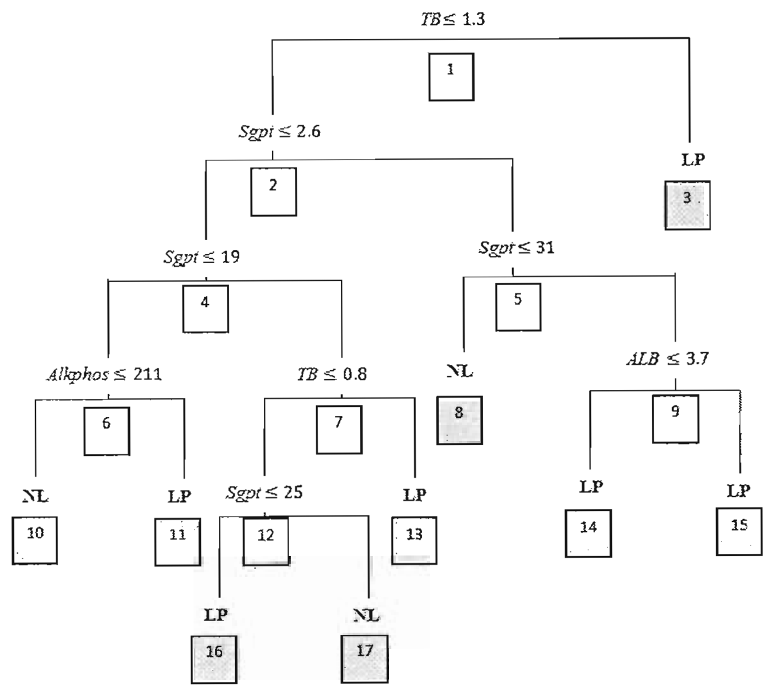


Figure 4.58. Winsorize tree of ILPD

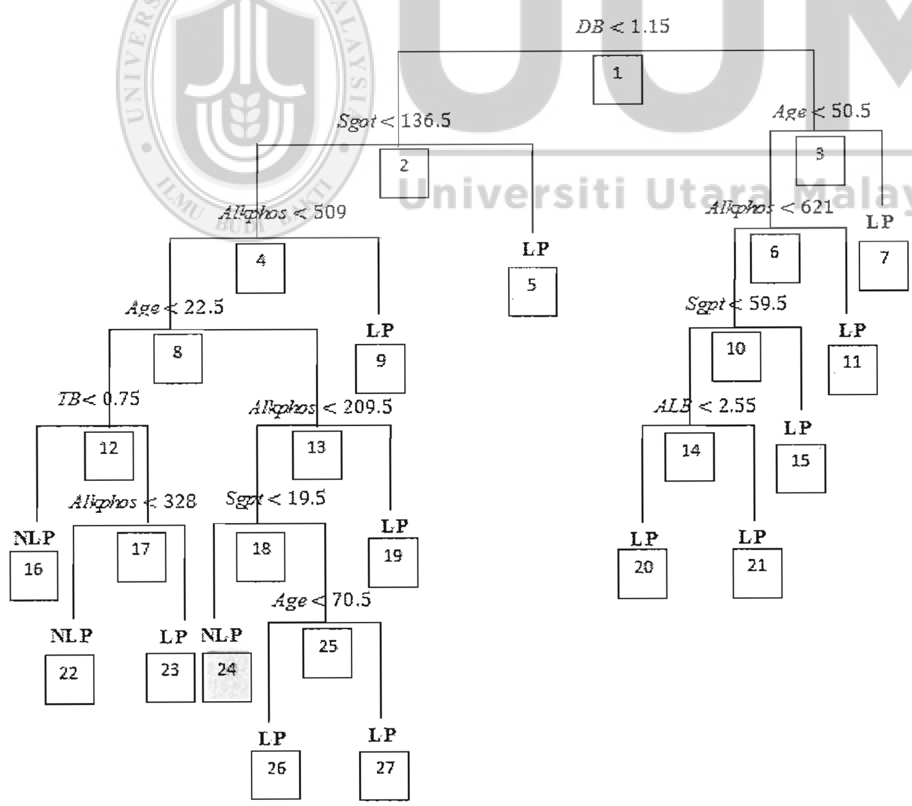


Figure 4.59. Traditional tree of ILPD

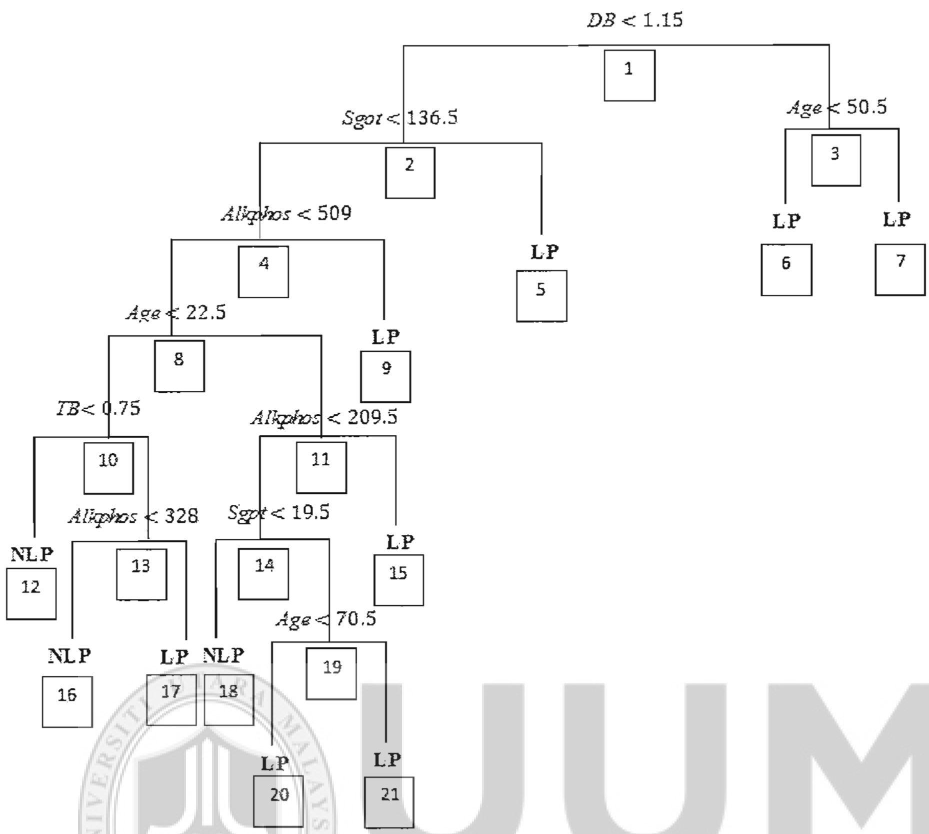


Figure 4.60. Pruned tree of ILPD

4.8.3 The Evaluation of Winsorize Tree for ILPD

To evaluate the performance of Winsorize tree, all trees are compared to determine whether Winsorize tree is able to compete with the traditional tree.

Table 4.68

Comparison between Traditional Tree, Pruned Tree and Winsorize Tree

ILPD:	Traditional Tree	Pruned Tree	Winsorize Tree
i. Number of splitting	13	10	8
ii. Number of leaves	14	11	9

ILPD:	Traditional Tree	Pruned Tree	Winsorize Tree
iii. Number of variable use	7	6	4
iv. Name of variable used	1. Age 2. DB 3. Sgpt 4. Sgot 5. Alkphos 6. TB 7. ALB	1. Age 2. DB 3. Sgpt 4. Sgot 5. Alkphos 6. TB	1. Age 2. TB 3. Alkphos 4. Sgpt
v. Error rate	**0.3316**	**0.3316**	***0.3109**
vi. Extreme value detected:			
a. First node	-	-	267
b. Second node	-	-	170
c. Forth node	-	-	68
d. Fifth node	-	-	103
e. Sixth node	-	-	16
f. Seventh node	-	-	42
g. ninth node	-	-	35
h. twelfth node	-	-	12

According to the result, Winsorize tree is having the least split which is only 8 splits compared to traditional tree and pruned tree which are 13 splits and 10 splits respectively. Besides, the variables used are fewer in Winsorize tree compared to the other trees. Only 4 variables are chosen during the construction of tree which these 4 variables are able to produce a better tree with lower error rate and simpler tree. In addition, all outliers are screened from level to level to make sure the data is in the accepted fence. In short, Winsorize tree performed even better in all forms compared to traditional tree and existing tree.

4.9 Case 7: Classification in Kyphosis Data

Kyphosis is called round back or Kelso's hunchback. This data contains 81 observations with 4 variables that representing the children who had corrective spinal surgery (Chamber & Hastie, 1992). In fact, this disease can be happened at any age even children. There are many factors that causing the curving of spine making the exaggerated rounding of the back.

Kyphosis data set includes 3 predictors which are `Age`, `Number` and `Start`. The target groups are whether "`absent`" or "`present`" indicate the type of deformation. According to information in rpart package in R, the variable `Age` is measured in months and variable `Number` represent the number of vertebrate involved. And the variable `Start` shows the number of the first vertebra operated on. The data has been split into training and test set where 54 observations are selected randomly to be the training set and the rest are used as test set. Figure 4.61 shows the picture of normal spine and kypho spine.

Figure 4.61(a). Normal spine

Figure 4.60(b). Kypho spine

4.9.1 The Statistical Background of the Kyphosis Data

The distribution of 54 training set is tabulated in Table 4.68. There are 43 from the group of absent and the rest from the group of present. Meanwhile, Table 4.69 summarises some descriptive statistics in order to give an overview about the behavior of each measured variables namely `Age`, `Number`, and `Start`. The standard deviation in variable of `Age` is extremely high and it may reflect the existence of outlier. However, the value of kurtosis and the value of skewness in `Age` and `Start` do not indicate any sign of having outlier as the values are in the range of [-2.00, 2.00]. In contra, the value of kurtosis in `Number` is slightly high, therefore we suspects that Number may have few outliers in it. In short, the empirical evidences of Kyphosis shows that the distribution of the data is quite symmetry. Further information is tabulated from Table 4.69 to Table 4.70.

Table 4.69

Frequency Table of Kyphosis Data Set

Class of Kyphosis	absent	present	Total
Frequency	43	11	54

Table 4.70

Statistical Description of Kyphosis Data Set

Variables	Mean	Median	Std. Deviation	Variance	Skewness	Kurtosis
Age	92.24	101.00	56.71	3216.45	-0.11	-1.05

Variables	Mean	Median	Std. Deviation	Variance	Skewness	Kurtosis
Number	4.13	4.00	1.71	2.91	1.33	2.345
Start	11.59	13.00	4.61	21.23	-1.02	0.19

Boxplot and bar chart are also used for further investigation on the existence of outlier and the attempt to highlight the separation between classes. Figure 4.62 and Figure 4.63 display the diagram of boxplot and bar chart respectively.

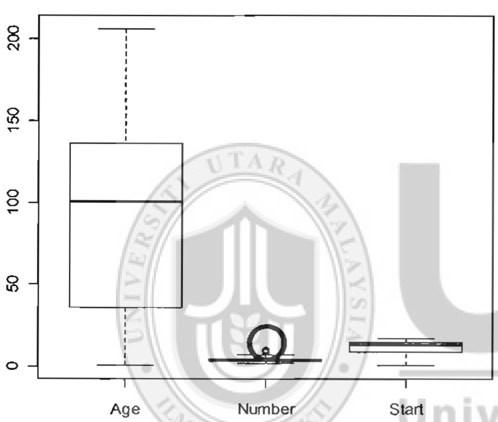


Figure 4.62. Outlier detection using boxplot

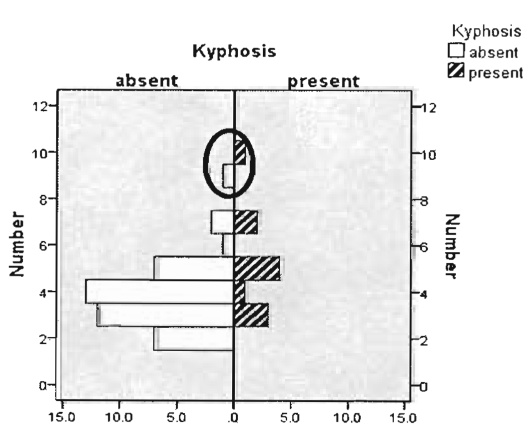


Figure 4.63(a). Original data of variable Number

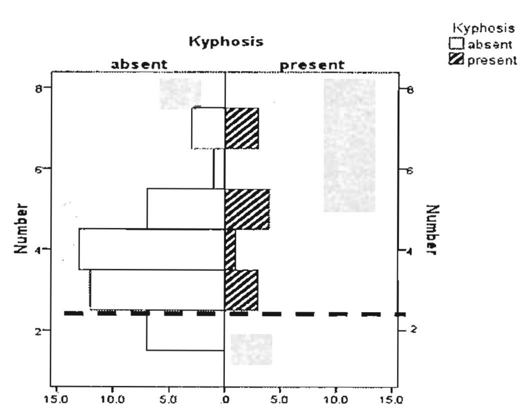


Figure 4.63(b). Winsorize data of variable Number

According to Figure 4.63(a), it is clear to see that the variable of `Number` contain outliers. Therefore, Winsorize need to be carried out to neutralise those outliers.

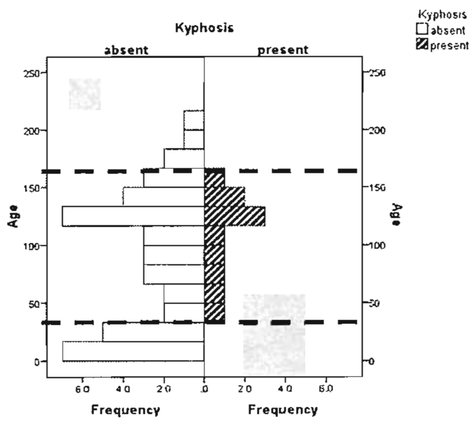


Figure 4.64. Original data of variable Age

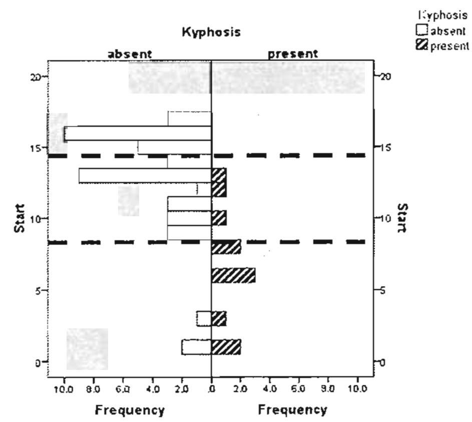


Figure 4.65. Original data of variable Start

According to Figure 4.63(b), Figure 4.64 and Figure 4.65, there are few possible potential cutting points. However, searching for the highest maximum homogeneity of group is put as priority. In variable Age, there are two clear splitting points which are about 175 or 40. Conversely, the possible splitting points of variable Start are about 8.5 or 14.0. The splitting point of variable of `Number` is unclear as both groups are overlapping to each other. The only clearest splitting point is about 2.5.

Table 4.71

Normality Tests

Variables	**Kolmogorov-Smirnov**			**Shapiro-Wilk**		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	0.12	54	**0.05**	0.95	54	**0.02**
Number	0.20	54	**0.00**	0.87	54	**0.00**
Start	0.19	54	**0.00**	0.87	54	**0.00**

Based on the normality test in Table 4.70, both `Number` and `Start` are not normally distributed as the p-value is less than 0.05. However, `Age` is approximately normal in Kolmogorov-Smirnov test as the value is exactly 0.05.

4.9.2 The Construction of Winsorize Tree for Kyphosis Data

The boxplot is capable to identify some outliers from each variable of the Kyphosis data (see Table 4.72).

Table 4.72

Outliers in Parent Node

Variable	Age	Number	Start
Number of outliers	0	2	0

Table 4.72 shows that only 2 outliers are found in variable `Number`. The suspicious values have been winsorized at 10% before computing the Gini purity index to determine the most potential variable to be used as a split variable in the parent node. Among these variables, `Start` with the splitting point of 8 gives the highest weighted average hence it is chosen in the first split. The table of Gini purity index is showed as in Table 4.73

Table 4.73

Splitting Point in Parent Node

Variable	Age	Number	Start
Highest weighted average	0.7046	0.7098	0.8158
Location of split	13th	3th	4th SP: 8

For the splitting process, those observations with the *Start* less than or equal to 8 will be assigned to the left node, t_l, and the remaining observations will be assigned to the right node, t_r. There are 11 observations and 43 observations of the original data are split into left (node 2) and right node (node 3) respectively as shown in Figure 4.66.

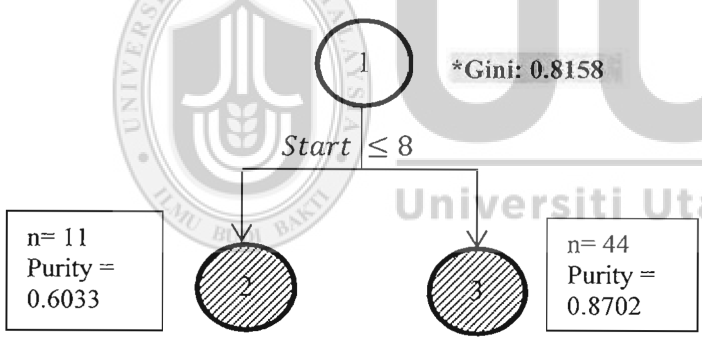


Figure 4.66. Child nodes from node 1

Table 4.74.

Number of Observations in Node 2 and Node 3

Node \ Group	absent	present
Node 2	3	8
Node 3	40	3

Due to the Gini purity index within variable in node 1 has already achieved the threshold (> 0.7), the node is allowed to be split into the final nodes (node 2 and node 3). In node 2, there are 3 in the group of absent and 8 in the group of present. And, in node 3, there are 40 objects and only 3 objects in the group of absent and present respectively. In short, node 3 is much pure than node 2 as it has achieved its maximum homogeneity. The final structure of traditional tree, pruned tree and Winsorize tree are shown in the following Figures.

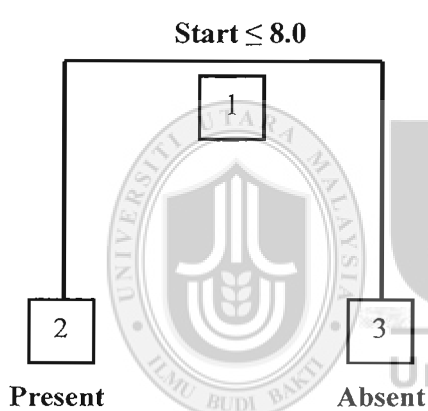


Figure 4.67. Winsorize tree of Kyphosis

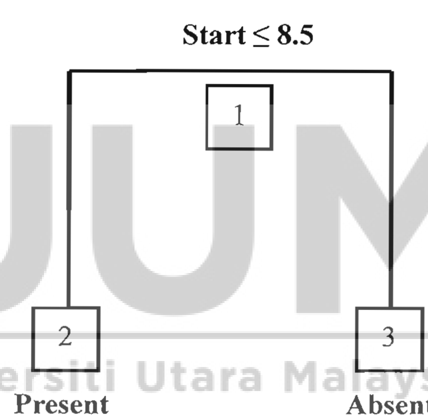


Figure 4.68. Traditional tree of Kyphosis

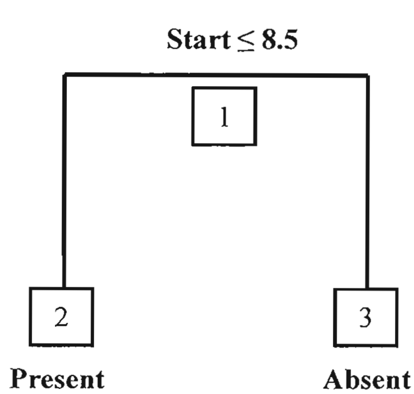


Figure 6.69. Pruned tree of Kyphosis

In this case, the pruned tree cannot be pruned as the original tree has only 2 terminal nodes. Therefore, both tree produced the same result of tree. Winsorize tree is also similar to traditional tree; the only different is the splitting point which is 8.0 compared to traditional tree (8.5).

4.9.3 The Evaluation of Winsorize Tree for Kyphosis Data

After the completion of the trees, we examined each tree to check for their similarities and differences. All structured of trees are recorded and tabulated in Table 4.75.

Table 4.75

Comparison between Traditional Tree, Pruned and Winsorize Tree

KYPHOSIS:	**Traditional Tree**	**Pruned Tree**	**Winsorize Tree**
i. Number of splitting	1	1	1
ii. Number of leaves	2	2	2
iii. Number of variable use	1	1	1
iv. Name of variable used	1. Start	1. Start	1. Start
v. Error rate	**0.2963**	**0.2963**	**0.2963**
vi. Extreme value detected:			
a. First node	-	-	2

Table 4.75 showed the result of traditional tree, pruned tree, and Winsorize tree. All trees have the same structure of tree with similar splitting point from variable of `Start`. Due to the small data set, only one split produced with final error rate of 0.2963. However, 2 extreme values have been detected which are from variable of `Number`. Even though there is no difference between these three types of tree, but at

least we know that the Winsorize tree is produced as good as traditional tree when dealing with small data sets. Moreover, it is capable to spot and to tolerate with outliers though the size of data is limited.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Introduction

Based on the results presented extensively in Chapter 4, our proposed algorithm has been proven workable and comparable to the traditional trees. In various fields, our results are always showing as good as the traditional tree or even better with the balance size. We summarise the investigation on the seven data sets as previously discussed in chapter 4 in Table 5.1.

Table 5.1

Overall Results of Seven Cases

Data size		Small		Medium			Big	
Cases		Case 5	Case 7	Case 1	Case 2	Case 4	Case 6	Case 3
Area		Life	Medicine	Medicine	Archeology	Life	Medicine	Medicine
Total Observations		49	81	106	150	150	583	768
Data Name		**Bumpus**	**Kyphosis**	**Breast Tissue**	**Egyptians skull**	**Iris**	**ILPD**	**Pima Indians**
Error rate	Winsorized tree	*0.5625	*0.2963	*0.2038	*0.7568	*0.06	*0.3109	*0.1758
	Traditional tree	0.6875	*0.2963	0.3846	0.8108	*0.06	0.3316	0.2188
	Pruned tree	0.6875	*0.2963	0.4231	*0.7568	*0.06	0.3316	0.2656
Number of splitting	Winsorized tree	5	1	7	11	2	8	8
	Traditional tree	3	1	7	14	2	13	13
	Pruned tree	3	1	6	9	2	10	8
Number of leaves	Winsorized tree	6	2	5	12	3	9	9
	Traditional tree	4	2	6	15	3	14	14
	Pruned tree	4	2	5	10	3	11	9
Number of variable used	Winsorized tree	4	1	5	4	1	4	5
	Traditional tree	3	1	6	4	1	7	8
	Pruned tree	3	1	5	4	1	6	5
Outliers detected	Winsorized tree	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Traditional tree	No	No	No	No	No	No	No
	Pruned tree	No	No	No	No	No	No	No

From the table presented, no matter how small or how big the data is, Winsorize tree is always producing a balance size and low error rate. We gained the lowest error rate in case 1, case 2, case 3, case 5 and case 6; we also gained the same result compared to the traditional and pruned trees in case 4 and case 7. In term of tree size, Winsorize tree showed a comparable or even smaller size compared to the traditional tree. However, the size is slightly bigger in case 1, case 2 and case 5 compared to the pruned tree.

Overall, Winsorize tree is more reliable as all the outliers are inspected and penalised in every single node. In the nutshell, taking good care along the process is vital in order to gain a more accurate and balance size of tree.

5.2 Achievement of Stopping Rules

Constructing a bushy or overfitting tree is sometime unrealistic. It causes time consuming. Moreover, the practitioner needs to do double tasks: constructing the tree and pruning the tree. In Winsorize tree, we introduce an easy stopping rule to assist the users to construct an acceptance tree. We set some thresholds in every single node so that the node can stop from continuing splitting once one of the thresholds is met. As mentioned in Chapter 3, the flexibility of thresholds are decided by the practitioner based on the background of the data or certain level of practitioners' knowledge.

There are three types of stopping rules to stop the tree from growing. Firstly, when the node achieved the minimum 10% of total training set, n_{min}. In case 1, for instance, node 14 (gla) is considered terminal node although the purity index is only as low as

0.2813 due to the minimum number (8) which is less than n_{min} (16). Secondly, if the node exceeds 0.7 of the overall Gini purity measurement then we considered the node as the terminal node. For example, since node 3 in case 1 gains the Gini purity index of 0.8892, we considered the purity of the node is sufficient enough to stop splitting. The last threshold achieved if one of the Winsorize Gini purity indexes between and within the variables is more than 0.7 during the variable splitting selection, the node will be split for the final nodes. The phenomenon is shown in case 3 (node 3), case 4(node 3), case 5 (node 3, node 4, and node 5) and case 7 (node 1).

Overall, if the data is having a very clear cutting point to separate the groups, the first and the third stopping rules are normally workable. However, if the data is complicated which all the groups are swamped together, normally second stopping rule is used to stop the tree. To achieve the minimum number of observations n_{min} is not an easy task especially for big or complicated data, therefore, we expect that the final tree in this case would be bushy.

5.3 Conclusion of Study

Classification tree has been widely studied for more than three decades for various aims. As part of contribution to the continuous development on this tool for classification, this study has focused on developing a tree which is insensitive towards the occurrence of outliers using Winsorize method. The idea of developing this tree is to replace a common strategy in handling bad data. Often, one has to validate a data prior to the construction of a classifier, which a strategy that best used by experts. Or, pruning process is implemented after the construction of tree is done. Thus, the proposed work embedded the strategy of handling outliers during the construction of a tree in an attempt to assist practitioners in general fields of studies.

The primary objectives of this study are: (i) to determine outlier in a data prior to construct the branch of tree, (ii) to manage the identified outliers accordingly using Winsorize method, (iii) to integrate the process of determining outlier and identifying outliers with the recursive process of constructing a tree and (iv) to propose Winsorize stopping criteria in constructing tree in order to avoid an over-fitting tree. In order to understand whether the proposed Winsorize tree is difference to traditional tree, some comparisons were made.

The stage of pre-constructing tree is vital such that all the data has to be screened and investigated to detect possible outliers. Each variable has to go through the outliers identification process by using boxplot. Boxplot is a simple yet powerful tool that has been widely used in exploratory data analysis. Any value falls outside the bound of the tolerance range, $[Q \pm 1.5 \times IQR]$ will be classified as extreme value or outliers. It can be used to detect even an individual outlying data point. In our study, detecting each potential outlier is vital to us so that we know which objects are significantly distorted. Then, the values need to be neutralised by Winsorize method to minimise the variability in Gini purity measurement. Based on the results that we have gained from the seven investigated cases, Winsorized tree is comparable to traditional tree, and sometimes even better. The Winsorize tree produced a simpler tree which insensitive variables are excluded. Moreover, since outliers are handled in every node, the final tree does not require pruning process.

According to the results presented in chapter 4, Winsorize tree is much precise and finally produces a high quality and accurate tree. The recorded error rate of Winsorize tree is lower compared to the traditional tree. The structure of Winsorize tree might be smaller but reliable as all splits are based on Gini purity index where all the contaminated data have been handled before the measurement. All the data are protected as no data is terminated. And, once the Gini purity index measurement is computed in a node, the original data will be reused for the following nodes. In short, the initial behavior of the data is in fact remains unchanged from the beginning till the end of the process. Since deep care has been implemented in all nodes, pruning process can be excluded once the Gini purity index has achieved the threshold (Gini purity index is more than 0.7).

In this study, three thresholds in stopping criteria have been set from creating a bushy tree. The aim is for time saving by the practitioners with a reliable tree classifier with pruning process is avoided. First, when the node contains 70% or above of homogeneity then the node will stop spitting. Secondly, when the node meets the minimum observation, n_{min} , which being set as to have 10% or 15% of total observations, n. However, this is depending on the practitioners' requirement. The lower the percentage of observation set, the bushy the tree it will be. Last but not least, computed *Gini* index between and within variables is greater than 70% or higher is absolutely vital as this process indicated whether the tree should stop before overfitting. Taking good care in this process can avoid a complicated tree. Therefore, cost complexity pruning process can be excluded. The findings given in Chapter 4 showed that the proposed method produces a simpler structure of tree with high accuracy output. In short, the new proposed method is comparative to the existing one or even better.

Throughout the thesis, we provide a better way in constructing a tree classifier especially in dealing with data which contains outliers. All outliers are investigated and handled during the process on creating a new binary branch. Thus, the structure of tree and the outcomes are strongly reliable. This phenomenon could bring an alternative way in classification for data mining. This method could be another potential tool in tree classification when the data contains outliers.

5.4 Contribution of Study

The practitioners do not have to pre-process the data, they proceed with the classification. Our proposed algorithm allows the practitioners to process during the construction of tree. Moreover, in real life, not all the practitioners are experts in dealing with the data. Or maybe the practitioners do not have time to go through all the historical data. Therefore, what they need is a trustable and reliable method with the method itself could be able to resist the abnormal data set and protecting the original information of the data. This study produces Winsorize tree algorithm which provides simultaneous data cleaning and model construction using Winsorize method along the tree growing process. It automatically investigates, detects, penalises and accommodates the suspicious value in all nodes to reduce the effect of contaminated data before performing Gini purity measurement. The Winsorize Gini purity index gained is able to resist to outliers while performing data splitting process. Besides, the proposed stopping criteria are able to stop the tree at the right time with the right size. Therefore, pruning process is not required in this study. In the nutshell, Winsorized tree algorithm is capable to produce a comparable or even better tree called Winsorize tree with no data are excluded along the construction of tree.

5.5 Limitation

However, this work has not yet answered the entire problem due to its limitation. For instance, outlier detection is limited to continuous variables instead of mixed variables. Also, this method is implemented to classification tree which means that the target group is categorical.

5.6 Future Works

Therefore, future work is necessary to fill on some gaps so that the tree can be widely applied in all fields such as marketing segmentation, banking loan credibility, risk analysis, logistic, supply chain management, medical diagnostic, sales analysis and so forth. Extending to this study, we may try on a huge, massive and more complex data set in future. In addition, dealing with missing value is another challenge that we should pay the focus on. Perhaps more questions may arise from real problem; therefore more studies on the application should be made to refine the method from time to time.

REFERENCES

Abraham, B., & Ledolter, J. (2006). *Introduction to regression modeling.* Belmont, USA: Thomson Higher Education.

Acuna, E., & Rodriguez, C. A. (2004). Meta analysis study of outlier detection methods in classification, Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Retrieved from academic.uprm.edu/ eacuna/ paperout.pdf. In proceedings IPSI 2004, Venice, 2004.

Altman, D. G., & Bland, J. M. (2009). Parametric *v* non-parametric methods for data analysis. *BMJ* 2009;338:a3167

Apte, C., & Weiss, S. (1997). Data mining with decision tress and decision rules. *Future Generation Computer Systems, 13*, 197–210.

Baesens, B., Van Gestel, T., Viaena, S., Stepanova, M., Suykens, J., & Vanthlenen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research, 54*(6), 249-268.

Bahrololum, M., & Khaleghi, M. (2008). Anomaly intrusion detection system using hierarchical gaussian mixture model. *Journal of Computer Science and Network Security, 8*(8), 264-271.

Barnett, V. (1978). The study of outliers: purpose and model. *Journal of Apllied Statistics, 27*(3), 242-250.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data.* (2nd ed.). New York: John Wiley.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: John Wiley.

Becker, R. A., Cleveland, W. S., & Wilk, A. R. (1987). Dynamic graphics for data analysis. *Journal of Statistical Science, 2*(4), 355-383.

Beguin, C. & Hulliger, B. (2004). Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. *Journal of the royal statistical society, 167*(2), 275-294.

Bensen, B., Gestel, T. V., Stepanova, M., Van den Poel, D., & Vanthienen, J. (1995). Neural network survival analysis for personal data. *Journal of the Operational Research Society, 56*(9), 1089-1098.

Bertolini, M., & Bevilacqua, M. (2006). Methodology and theory oil pipeline spill cause analysis: a classification tree approach. *Journal of Quality in Maintenance Engineering, 12*(2), 186-198.

Ben-Gal, I. (2005). *Outlier detection.* US: Springer.

Bluman, A. G (2004). *Elementary statistics.* (2^{nd} ed.). New York: McGraw Hill.

Bolton, R. J. & Hand, D. J. (2002). Statistical fraud detection: a review. *Journal of Statistical Science, 17*(3), 235-249.

Bratko, I., & Bohanec, M. (1994). Trading accuracy for simplicity in decision trees. *Machine Learning, 15*(3), 223-250.

Bramer, M. (2013). *Principle of data mining*. Springer-Verlag London

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, Calif., U.S.A.: Wadsworth, Inc.

Breimen, L. (1996). Some properties of splitting criteria. *Machine Learning*, *24*(1), 41-47.

Bridge, P. D. & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank sum test in small samples applied research. *Journal of Clinical Epidemiology*, *52*(3), 229–235.

Chamber, R., Hentges, A., & Zhao , X. Q. (2004). Robust automatic methods for outlier and error detection. *Journal of Royal Statistical Society, 167*(2), 323-339.

Chaovalit, P., & Zhou, L. (2005). A comparison between supervised and unsupervised classification approaches. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (pp. 1-9), Hawaii: IEEE.

Cernick, M. R. (2008). *Bootstrap methods a practitioner's guide*. New York: John Wiley.

Christina, M. R. K. (2009). Nonparametric vs Parametric Tests of Location in Biomedical Research. *Amrican Journal of Ophthamology.147*(4), 571-572.

Chambers, J. M., & Hastie, T. J. (1992). *Statistical models in S.* Wadsworth and Brooks/Cole, Pacific Grove: CA.

Coles, S., & Rowley, J. (1995). Revisiting decision trees. *Journal of Management Decision, 33*(8), 46-50.

Cunning, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In P. Cunning & M. Cord (Eds). *Machine learning techniques for multimedia*. Springer.

Curnow, R. N., & Franklin, M. F. (1973). Some further problem in the classification of human chromosomes. *International Bimetric Society, 29*(3), 429-440.

Davies, L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statiscal Association, 88*(423), 782-792.

De'ath, G., & Fabricius, K. E. (2000). A powerful yet simple technique for ecological data analysis. *Journal of Ecology Society of America, 81*(11), 3178-3192.

De Veaux, R. D., & Hand., D. J. (2005). How to lie with bad data. *Journal of Statistical Science, 20*(3), 231-238.

Dixon, W. J. (1960). Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*. *31*(2), 385-391.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: John Wiley and sons.

Duda, R. O., Hart, P. E., & Stork, D. R. (2001). *Pattern Recognition.* The University of Michigan: Wiley.

Dunham, M. H. (2003). *Data mining introductory and advanced topics*. New Jersey: Prentice Hall.

Efron, B. (1983). Estimating the misclassification rate of a prediction rule: improvement cross validation. *Journal of the American Statistical Asscociation, 78*(382), 316-331.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstapping.* London: Chapman & Hall.

Engels, R. (1996). Planning tasks for knowledge discovery in databases; performing task-oriented user-guidance. *Proceedings of the 2nd int. Conf. on Knowledge Discovery in Databases* (pp 170-175). AAAI press.

Engels, R., Evans, B., Herrmann, J. & Verdenius, F. (Eds.) (1997). Proceedings of the workshop on Machine Learning Application in the real world; Methodological Aspects and Implications. *14th International Conference on Machine Learning.*

Engels, R., & Theusinger, C. (1998) Using a data metric for preprocessing advice for data mining applications. In Prade, H. (ed.). Proceeding of 13th European Conference on Artificial Intelligence (pp 430-434). John Wiley & Sons, Chichester.

Evans, V. P. (1999). *Strategy for detecting outliers in regression analysis: an introductory primer* (Report No. TM029440 ED427059). San Antonio: Texas A & M University.

Eygptian Skull Department. (n.d.) *The data and story library.* Retrieved from http://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html

Fawagreh, K., & Gaber, M. M., & Elyan, E. (2015). CoRR abs/1503.04996

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. 179-188.

Frank, E. (2000), *Pruning decision trees and lists* (Doctoral dissertation). Retrieved from http://www.cs.waikato.ac.nz/~eibe/pubs/thesis.final.pdf.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *SIGKDD Explorations Newsletter*. *15*(1). 1-9

Gentleman, J. F. & Wilk, M. B. (1975). Detecting outliers. II. Supplementing the direct analysis of residuals. *Journal of Biometrics, 31*(2), 387-410.

Geisser, S. (1975). The predictive sample reused method with applications. *Journal of American Statistical Association, 70*(350), 243-250.

Ghahramani, Z. (2004). Unsupervised learning. In *Bousquet, O., von Luxburg, U. and Raetsch, G. Advanced Lectures in Machine Learning.* (pp.72-112). Berlin: Springer-Verlag.

Goutte, C. (1997). Note on tree lunches and cross validation. *Neural Computational*, *9*(6), 1211-1215.

Groβ, J. (2003). *Linear regression analysis.* New York: Springer.

Grubbs, F. E. (1950). Sample criteria for testing outlying observation observations. *Annals of Mathematical Statistics.21*(1), 27-58.

Gupta, G. K. (2006). *Introduction to data mining with case studies.* New Delhi: Prentice Hall.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, 54*(3), 761-771.

Hadi. A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of Royal Statistical Society, 56*(2), 393-396.

Hadi, A. S & Simonoff, J. S. (1993). Procedure for the identification of multiple outliers in linear models. *Journal of American Statistical Association, 88*(424), 1264-1272.

Hair, J. F., Anderson, R., Tatham, R. L., Black, W. C. (1992). *Multivariate data analysis with reading.* (3rd ed.). New York: Macmillan

Hamilton, L.C. (1992). *Regressions with graphics: A second course in applied statistics*. Monterey, CA: Brooks/Cole.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of American Statistic Association, 69*(346), 383-393.

Han, J. & Kamber, M (2006). *Data mining.* Amsterdam: Elsevier.

Hand D.J. (1997). *Construction and assessment of classification rules*, University of Michigan: Wiley.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994). *A small handbook of small data*. London: Chapman & Hall

Haslett, J., Bradley, R., Craig, P., Unwin, A. & Wills, G. (1991). Dynamic graphics for exploring spatial data with applications to locating global and local anomalies. *The American Statistician*, *45*(3), 234–242.

Haughton, D & Oulabi, S. (1997). Direct marketing modeling with CART and CHAID. *Journal of Interactive Marketing, 11*(4), 42-52.

Hauskrecht,et al. (2010). Conditional outlier detection for clinical alerting. *AMIA Annual Symposium Proceeding* (pp. 286-290).

Hawkins, D. M. (1980). *Identification of outliers.* New York: Chapman and Hall.

Hildebrand, D. K. (1986). *Statistical thinking for behavioral scientists*. Boston: Duxbury.

Ho, T. J. (2004). *Data mining and data warehousing.* Singapore: Prentice Hall.

Hollander, M. & Wolfe, D. A. (1999). *Nonparametric statistical methods.* (2nd ed.). New York: John Wiley.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression.* (2nd ed.). Canada: John Wiley & Son.

Iglewicz, Boris & Hoaglin, D. C. (1993). *How to detect and handle outliers (volume 16).* Milwaukee, Wisconsin: ASQC.

Jacobs, R. (2001). *Outliers in statistical analysis: basic methods of detection and accommodation.* (Report No. TM032341 ED450151). San Antonio: Texas A & M University.

Jiang Wen yu & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistic in Medicine*, *26*(29), 5320-5334.

Joachims, T. (2005). *Text categorization with support vector machines: learning with many relevant features*. Germany: Springer Berlin Heidelberg.

John, G. H. (1995). Robust decision trees: removing outliers from databases. *KDD-95 Proceeding (pp.* 174-179), Menlo Park, CA,: AAAI.

Johnson, D. E. (1998). *Applied multivariate method for data analysis*. California: Duxbury Press.

Jossinet, J. (1996).Variability of impedivity in normal and pathological breast tissue. *Med. & Biol. Eng. & Comput*, *34*, 346-350.

Kantardzic, M. (2011). Data mining concepts, models, and algorithms (2nd ed.). Hoboken, New Jersey: Wiley.

Kardi, T. (2006). What is bootstrap sampling. Retrieved May 6, 2009, from http://people.revoledu.com/kardi/tutorial/Bootstrap/bootstrap.htm.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, *29* (2), 119–127.

Kaufman, L. & Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley: New York.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (2000). *To err is human: building a safer health system*. Washington: National Academy Press.

Kotsiantis, S.B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*(3), 249-218.

Koufakou, A., Secretan, J., Reeder, J., Cardona, K., & Georgiopoulos, M. (2008). Fast parallel outlier detection for categorical datasets using mapreduce. *International Join Conference on Neural Networks (IJCNN 2008).* 3298-3304, 2008.

Kyung, H. O., June, S. S., Doo, H. H., & Nam, S. K. (2011). Decision tree-based clustering with outlier detection for HMM-based speech synthesis. *12th Annual Conference of the International Speech Communication Association* (pp. 101-104), Florence, Italy: ISCA.

Lachenbrunch, P. A. (1975). *Discriminant Analysis*. New York: Hafner Press.

Larson, R. & Farber, B. (2006). *Elementary statistics (picturing the world).* (3rd ed). New Jersey: Pearson Prentice Hall.

Lisboa, P. G. J. (1992). *Neural networks: current applications*. London: Chapman & Hall.

Loh, W. Y. (2011). Classification and regression tree. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1* ,14-23.

Mahat, N. I. (2006). *Some investigations in discriminant analysis with mixed variables.* Ph.D. thesis. Exeter University, UK.

Mclachlan, G. J. (2004). *Discriminant analysis and statiscal pattern recognition.* Canada: Jon Wiley & Sons.

Miller, T. W. (2005). *Data and text mining: a business application approach.* Upper Saddle River, New Jersey: Prentice hall.

Mingers, J. (1987). Expert systems—rule induction with statistical data. Journal of the Operational Research Society, 38, 39–47.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics, 21*(15), 3301-3307.

Muniyanni, A. P., Rajeswari, R., & Rajaram, R. (2011). Network anomaly detection by cascading K-means clustering and C4.5 decision tree algorithm. *Procedia Engineering,30* (2012), 174-182.

Newton, R.R., & Rudestam, K.E. (1999). *Your statistical consultant: Answers to your data analysis questions*. Thousand Oaks, CA: Sage.

Ng, R.T. & Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining, *In Proceedings of Very Large Data Bases Conference*, 144-155.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O Psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*(3), 473-486.

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation., 8*. Retrieved on April 5, 2014, from http://ericae.net/pare/getvn.asp?v=8&n=6.

Octavian (2011). Decision tree-C4.5 [Octavian's blog]. Retrieved Sept 10, 2014, from http://octaviansima.wordpress.com/2011/03/25/decision-trees-c4-5/

Parisot, O., Ghoniem, M., & Otjacques, B. (2014). Decision trees and data preprocessing to help clustering interpretation. *The 3rd International Conference on Data Management Technology and Applications (pp. 48-55). Vienna Austria.*

Penny, K. I. (1996). Appropriate critical values when testing for a single multivatiate outlier by using the mahalanobis distance. *Journal of Applied Statistics, 45*(1), 73-81.

Quenoullie, M. (1949). Approximate tests of correlation in time series. Journal of the Royal Statistical Society B, *11*, 18-44.

Quinlan, J. R. (1987). Simplying decision tree. *International Journal of Man-Machine Studies - Special Issue: Knowledge Acquisition for Knowledge-based Systems, 27*(3), 221-234.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. USA: Morgan Kaufmann Publishers.

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the Gini Index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77-93.

Rajendran, P., Madheswaran, M. & Naganandhini, K. (2010). An improved preprocessing technique with image mining approach for the medical image classification. *Second International Conference on Computing and Networking Technologies* (pp. 1-7).

Reif, J. M., Goldstein, M., Stahl. A., & Breuel, T. (2008). Anomaly detection by combining decision trees and parametric densities. *In ICPR 2008. IEEE*, 1-4.

Rokach, L., & Maimon, O. (2008). *Data Mining with decision trees theory and applications* (Vols. 69). Singapore: World Scientific.

Rousseeuw. P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection.* Hoboken, New Jersey: Wiley.

Sarkar, M., & Leong, T. Y. (2000). Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. *AMIA Annual Symposium Proceedings Archive* (pp. 759-763).

Sawilowsky, S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology. 52*(3). 229-235.

Seber, G. A. F. (1977). *Linear regression analysis.* Canada: John Wiley & Son.

Schurmann, J. (1996). *Pattern classification: A unified view of statistical and neural approaches.* New York: Wiley.

Shouman, M., Turner, T. & Stocker, R. (2011). Using Decision Tree for Diagnosing Heart Disease Patients. *In Proc. Australasian Data Mining Conference (AusDM 11) Ballarat, Australia. CRPIT* (pp. 23-29).

Silva, J. E., Marques de Sá, J. P., & Jossinet, J. (2000). Classification of Breast Tissue by Electrical Impedance Spectroscopy. *Med & Bio Eng & Computing, 38*, 26-30.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proc Annu Symp Comput Appl Med Care* (pp. 261-265). IEEE Computer Society Press.

Tabia, K. & Benferhat, S. (2008). On the use of decision trees as behavioral approaches in intrusion detection. *In Seventh International Conference on Machine Learning and Applications (ICMLA'08)*, IEEE, 665-670.

Terabe, M., Katai, O., Sawaragi, T, Washio. & Motoda, H. (1999). A data pre-processing method using association rules of attributes for improving decision tree. *Methodologies for Knowledge Discovery and Data Mining Lecture Notes in Computer Science*, 1574, 143-147.

Thomas, L. C., Oliver, R. W., & Hand, D. J. (2005). A survey of the issues in consumer credit modeling research. *Journal of the Operational Research Society, 56*(9), 1006-1015.

Timofeev, R. (2004). *Classification and regression trees (cart) theory and applications*. Master's thesis, Humboldt University Berlin.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology, 49*(11), 1225-1231.

Valera, V. A., Walter, B. A., Yokohama, N., Koyama, Y., Liai, T., & Okamoto. H. (2006). Prognostic groups in colorectal carcinoma patients base on tumor call proliferation and classification and regression. *Annals of Surgical Oncology, 14*(1), 34-40.

Wang, J. F., Gu, Y. S., & Wang, X. Z. (2004). Analysis of robustness about decision tree induced by insensible attribute. *Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai,* 1874-1877.

Wang, M. C. & Johnson, M. E. (n.d.). *Statistical decision theory in evaluating classification rules.* Retrieved from http:// pegasus.cc.ucf.edu/~cwang/sta6714/Lecture6/Note/Statistical%20Decision%20Theory.pdf.

Webb, A. (1999). *Statistical pattern recognition.* London: Arnold.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing.* San Diego, CA: Academic Press.

Wilkinson, L. (1992). Tree Structure data analysis: AID, CHAID and CART. *Paper presented at the 1992 Sun Valley, ID, Sawtooth/SYSTAT joint Software Conference.*

Wu, M.C., Lin, S.Y., & Lin, C.H. (2006). An effective application of decision tree to stock trading. *Expert Systems with Applications*, *31*(2), 270–274.

Xia, T., & Zhang, D. (2005). Improving the R*-tee with outlier handling techniques. *Proceeding of the Annual ACM International Workshop on Geographic Information Systems* (pp. 125-134). Bremen Germany: ACM

Xu, M., Wang, J. L., & Chen, T. (2006). Improved decision tree: ID3. In D. S. Huang, K. Li & W. Irwin. *Intelligent Computing in Signal Processing and Pattern Recognition* (pp.141-149). Berlin Heidelberg: Springer.

Young, F. M., Valero-Mora, P. M., & Friendly, M. (2006). *Visual statistics: seeing data with dynamic interactive graphics.* Hoboken, New Jersey: Wiley.

Zambon, M., Lawrence, R., Bunn, A., & Powell, S. (2006). Effect of alternative splitting rules on image processing using classification tree analysis. *American Society for Photogrammetry and Remote Sensing. 72*(1), 25-30.

Appendix A

Breast Tissue (Training and Test)

Training

Class	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
mas	172.52	0.13	0.04	37.54	192.22	5.12	19.32	32.19	174.93
con	650.00	0.04	0.15	216.81	427.53	1.97	33.77	214.17	528.70
mas	195.00	0.14	0.21	37.46	328.38	8.77	35.02	13.29	232.59
mas	544.65	0.06	0.00	100.79	1189.29	11.80	29.41	96.58	553.36
adi	2329.84	0.07	0.35	377.25	25369.04	67.25	336.08	171.39	2686.44
con	1461.75	0.04	0.05	391.85	5574.00	14.22	57.23	387.64	1428.84
con	1647.94	0.08	0.09	576.77	11852.49	20.55	111.44	565.90	1402.88
fad	272.00	0.09	0.00	63.79	718.95	11.27	20.09	60.69	286.92
con	1535.85	0.09	0.00	637.35	10814.05	16.97	96.61	632.17	1197.76
con	691.97	0.03	0.09	190.68	304.27	1.60	23.98	189.16	594.32
car	423.00	0.22	0.26	172.37	6108.11	35.44	79.06	153.17	558.27
adi	1850.00	0.07	0.23	325.19	8644.98	26.58	208.74	249.35	1908.18
car	500.00	0.23	0.05	219.30	9819.45	44.78	76.87	207.27	602.53
adi	2400.00	0.08	0.22	596.04	37939.26	63.65	261.35	535.69	2447.77
car	470.00	0.21	0.23	184.59	8185.36	44.34	84.48	164.12	603.32
car	438.78	0.21	0.06	120.90	4879.50	40.36	80.79	89.94	525.42
fad	200.00	0.04	0.12	42.32	220.81	5.22	10.68	40.95	218.03
con	649.37	0.11	0.02	207.11	3344.43	16.15	50.55	200.85	623.91
car	269.50	0.21	0.04	80.41	1963.61	24.42	44.74	66.84	329.09
adi	1700.00	0.04	0.11	120.65	12331.10	102.20	120.30	-9.26	2212.18
mas	260.28	0.08	0.03	58.82	277.26	4.71	17.87	56.04	248.62
gla	152.00	0.17	0.23	34.22	94.35	2.76	31.28	13.88	180.61
fad	211.00	0.05	0.09	30.75	151.98	4.94	14.27	27.24	217.13
mas	327.00	0.14	0.08	76.21	1664.67	21.84	43.22	62.77	379.26
gla	185.00	0.15	0.09	39.89	361.75	9.07	26.86	29.49	210.18
car	410.00	0.32	0.30	255.82	10622.55	41.52	67.52	246.74	508.54
gla	502.00	0.07	0.03	53.24	834.27	15.67	33.33	41.51	544.04
gla	250.00	0.09	0.09	29.64	180.76	6.10	26.14	13.96	280.12

mas	310.00	0.17	0.17	98.51	2741.03	27.82	49.33	85.27	388.98	
adi	1850.00	0.08	0.07	253.62	13113.20		51.70	160.07	196.73	1916.99
car	366.94	0.28	0.25	172.75	7064.82	40.90	75.60	155.32	471.59	
con	1500.00	0.06	0.05	375.10	4759.45	12.69	78.45	366.80	1336.16	
adi	2100.00	0.12	0.38	450.55	35671.61		79.17	436.10	113.20	2461.45
mas	370.40	0.10	0.00	115.92	1308.12	11.28	31.37	112.72	365.98	
car	330.00	0.23	0.27	121.15	3163.24	26.11	69.72	99.08	400.23	
fad	341.62	0.09	0.07	85.04	1370.84	16.12	29.03	79.94	385.13	
gla	216.41	0.12	0.07	53.60	280.45	5.23	22.79	48.51	215.37	
fad	196.86	0.02	0.09	28.59	82.06	2.87	7.97	27.66	200.75	
fad	155.00	0.17	0.12	38.94	415.11	10.66	25.84	29.13	184.82	
fad	352.66	0.12	0.09	68.53	1066.16	15.56	43.69	52.79	382.73	
car	300.00	0.19	0.17	97.11	3039.56	31.30	51.35	82.42	387.08	
adi	2600.00	0.07	0.05	745.47	39845.77		53.45	154.12	729.37	2545.42
adi	1600.00	0.07	-0.07	436.94	12655.34		28.96	103.73	432.13	1475.37
con	1111.81	0.10	0.07	386.99	7659.74	19.79	86.03	377.30	990.98	
mas	281.32	0.23	0.44	157.88	5305.12	33.60	46.38	150.92	398.90	
gla	197.00	0.13	0.07	33.46	409.65	12.24	26.99	19.77	231.78	
mas	250.00	0.05	0.01	70.91	224.15	3.16	9.10	70.32	232.28	
gla	178.00	0.15	0.10	40.29	474.40	11.77	25.92	30.85	209.18	
adi	1800.00	0.07	0.16	385.56	13831.72		35.87	157.57	351.90	1823.03
car	294.47	0.21	0.47	194.87	5541.26	28.44	36.77	191.80	445.51	
car	290.46	0.14	0.05	74.64	1189.55	15.94	35.70	65.54	330.27	
mas	435.09	0.08	0.16	123.60	1342.28	10.86	37.38	117.81	433.20	
con	1724.09	0.05	-0.02	404.13	3053.97	7.56	71.43	399.19	1489.39	
gla	124.13	0.13	0.11	20.59	78.34	3.80	18.46	9.12	134.89	
mas	274.99	0.15	0.14	66.46	1217.42	18.32	40.85	52.42	327.56	
car	390.00	0.36	0.20	245.69	10055.84		40.93	70.32	236.49	477.55
gla	303.00	0.06	0.04	22.57	102.50	4.54	21.83	5.72	321.65	
adi	2350.00	0.08	0.27	515.29	27758.64		53.87	289.57	426.23	2457.68
adi	1666.15	0.01	0.06	72.93	1402.23	19.23	51.85	58.60	1746.58	
mas	121.00	0.17	0.09	24.44	144.47	5.91	22.02	10.59	141.77	
adi	2000.00	0.11	0.11	520.22	40087.92		77.06	204.09	478.52	2088.65
gla	223.00	0.12	0.08	33.10	197.01	5.95	30.45	12.96	252.48	

gla	197.00	0.13	0.07	33.46	409.65	12.24	26.99	19.77	231.78	
car	325.00	0.22	0.29	229.22	5705.3324.89		35.60	227.26	462.70	
adi	1949.120.05		0.02	170.33	3212.0818.86		101.46	136.82	1941.37	
adi	2000.000.07		0.12	330.27	15381.10		46.57	169.20	283.64	2063.07
adi	1800.000.09		0.21	362.86	15021.55		41.40	217.83	290.20	1893.66
car	389.87	0.15	0.10	118.63	2475.5620.87		49.76	107.69	429.39	
adi	2600.000.20		0.21	1063.44174480.48			164.07	418.69	977.55	2664.58
mas	196.36	0.18	0.14	54.58	843.26	15.45	34.15	42.58	239.94	
car	500.00	0.19	0.19	144.69	3055.0121.11		96.56	107.75	542.90	
gla	176.00	0.09	0.08	20.59	79.71	3.87	18.23	9.58	191.99	
mas	236.00	0.12	0.20	48.45	236.88	4.89	36.01	32.42	244.97	
gla	103.00	0.16	0.29	23.75	78.26	3.29	22.32	8.12	124.98	
car	524.79	0.19	0.03	228.80	6843.6029.91		60.20	220.74	556.83	
fad	259.89	0.07	0.01	58.24	465.09	7.99	17.51	56.34	267.52	
mas	252.00	0.11	0.03	38.54	493.79	12.81	25.54	28.87	280.66	
fad	243.29	0.04	0.07	68.54	383.93	5.60	9.99	67.82	263.64	
gla	145.00	0.12	0.11	21.22	82.46	3.89	20.30	6.17	162.51	
con	1496.740.10		0.08	640.28	11072.00		17.29	108.29	631.05	1178.27

Test

fad	250.00	0.07	-0.02	57.17	652.90	11.42	17.78	55.79	278.31	
adi	2800.000.08		0.18	583.26	31388.65		53.82	298.58	501.04	2896.58
con	770.00	0.04	0.00	175.02	346.09	1.98	25.22	173.19	654.80	
fad	301.30	0.11	0.04	64.62	942.77	14.59	29.05	57.72	335.77	
adi	2100.000.06		-0.05	390.48	16640.72		42.62	125.90	380.64	2073.03
fad	245.00	0.19	0.08	62.90	1235.9819.65		42.15	46.69	292.38	
mas	339.51	0.05	0.03	88.63	331.08	3.74	19.83	87.62	307.79	
car	362.83	0.20	0.24	124.91	3290.4626.34		69.39	103.87	424.80	
con	1270.670.08		0.07	555.35	3612.976.51		68.78	551.08	895.19	
con	1385.660.09		0.09	202.48	8785.0343.39		143.09	143.26	1524.61	
fad	160.32	0.18	0.16	37.22	341.88	9.19	30.89	20.76	187.57	
car	485.67	0.23	0.13	253.89	8135.9732.04		64.86	245.47	541.36	
car	275.68	0.15	0.19	91.53	1756.2319.19		39.31	82.66	331.59	
con	1084.250.07		0.00	191.90	2937.9715.31		66.56	179.98	1064.10	

fad	144.00	0.12	0.05	19.65	70.43	3.58	18.13	7.57	160.37	
adi	1800.00	0.03	0.04	301.06	4406.15	14.64	67.63	293.37	1742.38	
gla	470.52	0.13	0.07	150.22	2657.91	17.69	47.56	142.50	491.47	
adi	1900.00	0.05	0.11	272.62	7481.59	27.44	138.36	234.90	1924.52	
adi	1650.00	0.05	0.04	274.43	5824.90	21.23	81.24	262.13	1603.07	
mas	178.00	0.17	0.21	41.54	489.44	11.78	35.75	21.16	215.91	
car	551.88	0.23	0.06	264.80	11888.39		44.89	77.79	253.79	656.77
fad	355.00	0.06	0.08	89.56	1033.85	11.54	27.56	86.58	372.04	
gla	391.00	0.06	0.01	35.78	265.15	7.41	22.13	28.11	400.99	
adi	2300.00	0.05	0.14	185.45	5086.29	27.43	178.69	49.59	2480.59	
car	380.00	0.24	0.29	137.64	5402.17	39.25	88.76	105.20	493.70	
mas	481.47	0.08	0.02	79.06	1154.34	14.60	33.93	71.41	501.89	

Appendix B

Egyptian Skulls (Training and Test)

Training

epoch mb	bh	bl	nh	
c200BC	141	128	95	53
c1850BC	138	137	94	51
c1850BC	129	135	92	50
c3300BC	130	129	105	47
c1850BC	130	127	99	45
c3300BC	131	128	98	45
cAD150	134	124	91	55
c1850BC	138	133	100	55
c200BC	138	140	100	52
c3300BC	132	130	104	50
c4000BC	128	134	103	50
c4000BC	136	143	100	54
c1850BC	130	134	106	50
c1850BC	133	131	96	49
c4000BC	131	134	102	51
c3300BC	135	132	98	54
c3300BC	135	136	98	52
c200BC	131	142	95	53
c3300BC	134	139	101	49
c200BC	134	137	93	52
c200BC	133	120	91	46
c3300BC	131	139	98	51
c3300BC	133	136	103	53
c4000BC	129	138	95	50
c200BC	129	135	95	47
c200BC	140	137	94	60
c1850BC	136	135	94	53
c1850BC	136	126	101	50

c4000BC	126	133	102	51
c1850BC	126	136	95	56
c200BC	134	134	97	54
c3300BC	131	136	99	56
c4000BC	125	131	92	48
cAD150	139	134	95	47
c3300BC	134	130	93	54
c1850BC	138	133	91	46
c200BC	132	133	90	53
c4000BC	138	135	100	55
c3300BC	131	134	96	50
c200BC	131	135	90	50
cAD150	136	138	97	58
c1850BC	137	141	96	52
c4000BC	124	138	101	46
c4000BC	138	137	89	56
c3300BC	126	124	95	45
c3300BC	148	129	104	51
c4000BC	132	133	93	53
cAD150	138	136	92	46
c4000BC	119	132	96	44
c200BC	141	130	87	49
c200BC	131	141	99	55
cAD150	147	129	87	48
c3300BC	137	136	106	49
c1850BC	138	138	95	47
c1850BC	140	133	98	50
c3300BC	130	136	104	53
cAD150	140	135	103	48
cAD150	137	123	91	50
c4000BC	135	135	103	47
c3300BC	130	132	93	52
c200BC	131	125	88	48
cAD150	141	136	101	54

c200BC	132	136	92	52
c4000BC	131	134	97	54
c4000BC	134	134	99	51
c200BC	139	130	94	53
c200BC	139	130	90	48
c1850BC	132	130	91	52
c4000BC	139	136	96	50
c4000BC	132	131	101	49
cAD150	137	134	93	53
c3300BC	138	134	98	49
c1850BC	137	133	90	49
c200BC	141	131	97	53
cAD150	133	125	92	50
c3300BC	130	128	101	51
c200BC	135	135	99	50
c3300BC	124	138	101	48
c4000BC	139	130	108	48
c3300BC	131	134	90	53
c200BC	137	134	107	54
c200BC	140	134	90	51
c3300BC	138	134	98	45
cAD150	138	125	99	51
c3300BC	133	130	102	48
c3300BC	129	126	91	50
c3300BC	133	134	97	48
c200BC	136	128	93	54
c4000BC	128	132	93	53
c4000BC	131	132	99	50
c1850BC	134	123	95	52
cAD150	132	127	97	52
c4000BC	141	140	100	51
cAD150	137	125	85	57
c3300BC	126	131	100	48
cAD150	129	128	81	52

c3300BC	132	145	100	54
cAD150	136	131	95	49
c1850BC	136	145	99	55
cAD150	136	133	97	51
cAD150	138	127	86	47
c1850BC	133	131	100	50
c1850BC	129	142	104	47
cAD150	143	120	95	51
cAD150	135	135	95	56
c200BC	133	128	92	51
c200BC	135	131	99	51
cAD150	132	132	99	55
c1850BC	138	134	96	51
cAD150	128	126	91	57
c1850BC	134	134	96	45
c4000BC	134	121	95	53
c4000BC	126	129	109	51

Test

epoch	mb	bh	bl	nh
c200BC	144	124	86	50
c1850BC	132	138	87	48
cAD150	143	126	88	54
c4000BC	131	138	89	49
cAD150	145	129	89	47
c1850BC	134	125	90	60
c1850BC	136	133	91	49
cAD150	126	126	92	45
cAD150	130	134	92	52
cAD150	139	135	92	54
c1850BC	136	131	92	46
c4000BC	125	136	93	48
c1850BC	129	133	93	47
c4000BC	134	124	93	53

c3300BC	133	125	94	46
c200BC	136	138	94	55
c200BC	133	136	95	52
cAD150	137	135	96	54
cAD150	142	135	96	52
c200BC	138	126	97	54
c1850BC	137	139	97	50
c3300BC	135	136	97	52
cAD150	131	129	97	44
c200BC	130	131	98	53
c200BC	136	130	99	55
c4000BC	132	136	100	50
c200BC	135	130	100	51
c1850BC	137	129	100	53
cAD150	126	138	101	52
c1850BC	136	137	101	54
c3300BC	123	131	101	51
c1850BC	135	138	102	55
c4000BC	135	137	103	50
c4000BC	130	130	104	49
c4000BC	127	129	106	48
c3300BC	138	129	107	53
c4000BC	131	136	114	54

Appendix C

Pima Indians (Training and Test)

Training

Num_preg	PGC	DBP	TRICEP	SERUM	BMI	DPF	AGE	CLASS
6	148	72	35	0	33.6	0.627	50	positive
1	85	66	29	0	26.6	0.351	31	negative
8	183	64	0	0	23.3	0.672	32	positive
1	89	66	23	94	28.1	0.167	21	negative
0	137	40	35	168	43.1	2.288	33	positive
5	116	74	0	0	25.6	0.201	30	negative
3	78	50	32	88	31	0.248	26	positive
10	115	0	0	0	35.3	0.134	29	negative
2	197	70	45	543	30.5	0.158	53	positive
8	125	96	0	0	0	0.232	54	positive
4	110	92	0	0	37.6	0.191	30	negative
10	168	74	0	0	38	0.537	34	positive
10	139	80	0	0	27.1	1.441	57	negative
1	189	60	23	846	30.1	0.398	59	positive
5	166	72	19	175	25.8	0.587	51	positive
7	100	0	0	0	30	0.484	32	positive
0	118	84	47	230	45.8	0.551	31	positive
7	107	74	0	0	29.6	0.254	31	positive
1	103	30	38	83	43.3	0.183	33	negative
1	115	70	30	96	34.6	0.529	32	positive
3	126	88	41	235	39.3	0.704	27	negative
8	99	84	0	0	35.4	0.388	50	negative
7	196	90	0	0	39.8	0.451	41	positive
9	119	80	35	0	29	0.263	29	positive
11	143	94	33	146	36.6	0.254	51	positive
10	125	70	26	115	31.1	0.205	41	positive
7	147	76	0	0	39.4	0.257	43	positive
1	97	66	15	140	23.2	0.487	22	negative

13	145	82	19	110	22.2	0.245	57	negative
5	117	92	0	0	34.1	0.337	38	negative
5	109	75	26	0	36	0.546	60	negative
3	158	76	36	245	31.6	0.851	28	positive
3	88	58	11	54	24.8	0.267	22	negative
6	92	92	0	0	19.9	0.188	28	negative
10	122	78	31	0	27.6	0.512	45	negative
4	103	60	33	192	24	0.966	33	negative
11	138	76	0	0	33.2	0.42	35	negative
9	102	76	37	0	32.9	0.665	46	positive
2	90	68	42	0	38.2	0.503	27	positive
4	111	72	47	207	37.1	1.39	56	positive
3	180	64	25	70	34	0.271	26	negative
7	133	84	0	0	40.2	0.696	37	negative
7	106	92	18	0	22.7	0.235	48	negative
9	171	110	24	240	45.4	0.721	54	positive
7	159	64	0	0	27.4	0.294	40	negative
0	180	66	39	0	42	1.893	25	positive
1	146	56	0	0	29.7	0.564	29	negative
2	71	70	27	0	28	0.586	22	negative
7	103	66	32	0	39.1	0.344	31	positive
7	105	0	0	0	0	0.305	24	negative
1	103	80	11	82	19.4	0.491	22	negative
1	101	50	15	36	24.2	0.526	26	negative
5	88	66	21	23	24.4	0.342	30	negative
8	176	90	34	300	33.7	0.467	58	positive
7	150	66	42	342	34.7	0.718	42	negative
1	73	50	10	0	23	0.248	21	negative
7	187	68	39	304	37.7	0.254	41	positive
0	100	88	60	110	46.8	0.962	31	negative
0	146	82	0	0	40.5	1.781	44	negative
0	105	64	41	142	41.5	0.173	22	negative
2	84	0	0	0	0	0.304	21	negative
8	133	72	0	0	32.9	0.27	39	positive

5	44	62	0	0	25	0.587	36	negative
2	141	58	34	128	25.4	0.699	24	negative
7	114	66	0	0	32.8	0.258	42	positive
5	99	74	27	0	29	0.203	32	negative
0	109	88	30	0	32.5	0.855	38	positive
2	109	92	0	0	42.7	0.845	54	negative
1	95	66	13	38	19.6	0.334	25	negative
4	146	85	27	100	28.9	0.189	27	negative
2	100	66	20	90	32.9	0.867	28	positive
5	139	64	35	140	28.6	0.411	26	negative
13	126	90	0	0	43.4	0.583	42	positive
4	129	86	20	270	35.1	0.231	23	negative
1	79	75	30	0	32	0.396	22	negative
1	0	48	20	0	24.7	0.14	22	negative
7	62	78	0	0	32.6	0.391	41	negative
5	95	72	33	0	37.7	0.37	27	negative
0	131	0	0	0	43.2	0.27	26	positive
2	112	66	22	0	25	0.307	24	negative
3	113	44	13	0	22.4	0.14	22	negative
2	74	0	0	0	0	0.102	22	negative
7	83	78	26	71	29.3	0.767	36	negative
0	101	65	28	0	24.6	0.237	22	negative
5	137	108	0	0	48.8	0.227	37	positive
2	110	74	29	125	32.4	0.698	27	negative
13	106	72	54	0	36.6	0.178	45	negative
2	100	68	25	71	38.5	0.324	26	negative
15	136	70	32	110	37.1	0.153	43	positive
1	107	68	19	0	26.5	0.165	24	negative
1	80	55	0	0	19.1	0.258	21	negative
4	123	80	15	176	32	0.443	34	negative
7	81	78	40	48	46.7	0.261	42	negative
4	134	72	0	0	23.8	0.277	60	positive
2	142	82	18	64	24.7	0.761	21	negative
6	144	72	27	228	33.9	0.255	40	negative

2	92	62	28	0	31.6	0.13	24	negative
1	71	48	18	76	20.4	0.323	22	negative
6	93	50	30	64	28.7	0.356	23	negative
1	122	90	51	220	49.7	0.325	31	positive
1	163	72	0	0	39	1.222	33	positive
1	151	60	0	0	26.1	0.179	22	negative
0	125	96	0	0	22.5	0.262	21	negative
1	81	72	18	40	26.6	0.283	24	negative
2	85	65	0	0	39.6	0.93	27	negative
1	126	56	29	152	28.7	0.801	21	negative
1	96	122	0	0	22.4	0.207	27	negative
4	144	58	28	140	29.5	0.287	37	negative
3	83	58	31	18	34.3	0.336	25	negative
0	95	85	25	36	37.4	0.247	24	positive
3	171	72	33	135	33.3	0.199	24	positive
8	155	62	26	495	34	0.543	46	positive
1	89	76	34	37	31.2	0.192	23	negative
4	76	62	0	0	34	0.391	25	negative
7	160	54	32	175	30.5	0.588	39	positive
4	146	92	0	0	31.2	0.539	61	positive
5	124	74	0	0	34	0.22	38	positive
5	78	48	0	0	33.7	0.654	25	negative
4	97	60	23	0	28.2	0.443	22	negative
4	99	76	15	51	23.2	0.223	21	negative
0	162	76	56	100	53.2	0.759	25	positive
6	111	64	39	0	34.2	0.26	24	negative
2	107	74	30	100	33.6	0.404	23	negative
5	132	80	0	0	26.8	0.186	69	negative
0	113	76	0	0	33.3	0.278	23	positive
1	88	30	42	99	55	0.496	26	positive
3	120	70	30	135	42.9	0.452	30	negative
1	118	58	36	94	33.3	0.261	23	negative
1	117	88	24	145	34.5	0.403	40	positive
0	105	84	0	0	27.9	0.741	62	positive

4	173	70	14	168	29.7	0.361	33	positive
9	122	56	0	0	33.3	1.114	33	positive
3	170	64	37	225	34.5	0.356	30	positive
8	84	74	31	0	38.3	0.457	39	negative
2	96	68	13	49	21.1	0.647	26	negative
2	125	60	20	140	33.8	0.088	31	negative
0	100	70	26	50	30.8	0.597	21	negative
0	93	60	25	92	28.7	0.532	22	negative
0	129	80	0	0	31.2	0.703	29	negative
5	105	72	29	325	36.9	0.159	28	negative
3	128	78	0	0	21.1	0.268	55	negative
5	106	82	30	0	39.5	0.286	38	negative
2	108	52	26	63	32.5	0.318	22	negative
10	108	66	0	0	32.4	0.272	42	positive
4	154	62	31	284	32.8	0.237	23	negative
0	102	75	23	0	0	0.572	21	negative
9	57	80	37	0	32.8	0.096	41	negative
2	106	64	35	119	30.5	1.4	34	negative
5	147	78	0	0	33.7	0.218	65	negative
2	90	70	17	0	27.3	0.085	22	negative
1	136	74	50	204	37.4	0.399	24	negative
4	114	65	0	0	21.9	0.432	37	negative
9	156	86	28	155	34.3	1.189	42	positive
1	153	82	42	485	40.6	0.687	23	negative
8	188	78	0	0	47.9	0.137	43	positive
7	152	88	44	0	50	0.337	36	positive
2	99	52	15	94	24.6	0.637	21	negative
1	109	56	21	135	25.2	0.833	23	negative
2	88	74	19	53	29	0.229	22	negative
17	163	72	41	114	40.9	0.817	47	positive
4	151	90	38	0	29.7	0.294	36	negative
7	102	74	40	105	37.2	0.204	45	negative
0	114	80	34	285	44.2	0.167	27	negative
2	100	64	23	0	29.7	0.368	21	negative

0	131	88	0	0	31.6	0.743	32	positive
6	104	74	18	156	29.9	0.722	41	positive
3	148	66	25	0	32.5	0.256	22	negative
4	120	68	0	0	29.6	0.709	34	negative
4	110	66	0	0	31.9	0.471	29	negative
3	111	90	12	78	28.4	0.495	29	negative
6	102	82	0	0	30.8	0.18	36	positive
6	134	70	23	130	35.4	0.542	29	positive
2	87	0	23	0	28.9	0.773	25	negative
1	79	60	42	48	43.5	0.678	23	negative
2	75	64	24	55	29.7	0.37	33	negative
8	179	72	42	130	32.7	0.719	36	positive
6	85	78	0	0	31.2	0.382	42	negative
0	129	110	46	130	67.1	0.319	26	positive
5	143	78	0	0	45	0.19	47	negative
5	130	82	0	0	39.1	0.956	37	positive
6	87	80	0	0	23.2	0.084	32	negative
0	119	64	18	92	34.9	0.725	23	negative
1	0	74	20	23	27.7	0.299	21	negative
5	73	60	0	0	26.8	0.268	27	negative
4	141	74	0	0	27.6	0.244	40	negative
7	194	68	28	0	35.9	0.745	41	positive
8	181	68	36	495	30.1	0.615	60	positive
1	128	98	41	58	32	1.321	33	positive
8	109	76	39	114	27.9	0.64	31	positive
5	139	80	35	160	31.6	0.361	25	positive
3	111	62	0	0	22.6	0.142	21	negative
9	123	70	44	94	33.1	0.374	40	negative
7	159	66	0	0	30.4	0.383	36	positive
11	135	0	0	0	52.3	0.578	40	positive
8	85	55	20	0	24.4	0.136	42	negative
5	158	84	41	210	39.4	0.395	29	positive
1	105	58	0	0	24.3	0.187	21	negative
3	107	62	13	48	22.9	0.678	23	positive

4	109	64	44	99	34.8	0.905	26	positive
4	148	60	27	318	30.9	0.15	29	positive
0	113	80	16	0	31	0.874	21	negative
1	138	82	0	0	40.1	0.236	28	negative
0	108	68	20	0	27.3	0.787	32	negative
2	99	70	16	44	20.4	0.235	27	negative
6	103	72	32	190	37.7	0.324	55	negative
5	111	72	28	0	23.9	0.407	27	negative
8	196	76	29	280	37.5	0.605	57	positive
5	162	104	0	0	37.7	0.151	52	positive
1	96	64	27	87	33.2	0.289	21	negative
7	184	84	33	0	35.5	0.355	41	positive
2	81	60	22	0	27.7	0.29	25	negative
0	147	85	54	0	42.8	0.375	24	negative
7	179	95	31	0	34.2	0.164	60	negative
0	140	65	26	130	42.6	0.431	24	positive
9	112	82	32	175	34.2	0.26	36	positive
12	151	70	40	271	41.8	0.742	38	positive
5	109	62	41	129	35.8	0.514	25	positive
6	125	68	30	120	30	0.464	32	negative
5	85	74	22	0	29	1.224	32	positive
5	112	66	0	0	37.8	0.261	41	positive
0	177	60	29	478	34.6	1.072	21	positive
2	158	90	0	0	31.6	0.805	66	positive
7	119	0	0	0	25.2	0.209	37	negative
7	142	60	33	190	28.8	0.687	61	negative
1	100	66	15	56	23.6	0.666	26	negative
1	87	78	27	32	34.6	0.101	22	negative
0	101	76	0	0	35.7	0.198	26	negative
3	162	52	38	0	37.2	0.652	24	positive
4	197	70	39	744	36.7	2.329	31	negative
0	117	80	31	53	45.2	0.089	24	negative
4	142	86	0	0	44	0.645	22	positive
6	134	80	37	370	46.2	0.238	46	positive

1	79	80	25	37	25.4	0.583	22	negative
4	122	68	0	0	35	0.394	29	negative
3	74	68	28	45	29.7	0.293	23	negative
4	171	72	0	0	43.6	0.479	26	positive
7	181	84	21	192	35.9	0.586	51	positive
0	179	90	27	0	44.1	0.686	23	positive
9	164	84	21	0	30.8	0.831	32	positive
0	104	76	0	0	18.4	0.582	27	negative
1	91	64	24	0	29.2	0.192	21	negative
4	91	70	32	88	33.1	0.446	22	negative
3	139	54	0	0	25.6	0.402	22	positive
6	119	50	22	176	27.1	1.318	33	positive
2	146	76	35	194	38.2	0.329	29	negative
9	184	85	15	0	30	1.213	49	positive
10	122	68	0	0	31.2	0.258	41	negative
0	165	90	33	680	52.3	0.427	23	negative
9	124	70	33	402	35.4	0.282	34	negative
1	111	86	19	0	30.1	0.143	23	negative
9	106	52	0	0	31.2	0.38	42	negative
2	129	84	0	0	28	0.284	27	negative
2	90	80	14	55	24.4	0.249	24	negative
0	86	68	32	0	35.8	0.238	25	negative
12	92	62	7	258	27.6	0.926	44	positive
1	113	64	35	0	33.6	0.543	21	positive
3	111	56	39	0	30.1	0.557	30	negative
2	114	68	22	0	28.7	0.092	25	negative
1	193	50	16	375	25.9	0.655	24	negative
11	155	76	28	150	33.3	1.353	51	positive
3	191	68	15	130	30.9	0.299	34	negative
3	141	0	0	0	30	0.761	27	positive
4	95	70	32	0	32.1	0.612	24	negative
3	142	80	15	0	32.4	0.2	63	negative
4	123	62	0	0	32	0.226	35	positive
5	96	74	18	67	33.6	0.997	43	negative

0	138	0	0	0	36.3	0.933	25	positive
2	128	64	42	0	40	1.101	24	negative
0	102	52	0	0	25.1	0.078	21	negative
2	146	0	0	0	27.5	0.24	28	positive
10	101	86	37	0	45.6	1.136	38	positive
2	108	62	32	56	25.2	0.128	21	negative
3	122	78	0	0	23	0.254	40	negative
1	71	78	50	45	33.2	0.422	21	negative
13	106	70	0	0	34.2	0.251	52	negative
2	100	70	52	57	40.5	0.677	25	negative
7	106	60	24	0	26.5	0.296	29	positive
0	104	64	23	116	27.8	0.454	23	negative
5	114	74	0	0	24.9	0.744	57	negative
2	108	62	10	278	25.3	0.881	22	negative
0	146	70	0	0	37.9	0.334	28	positive
10	129	76	28	122	35.9	0.28	39	negative
7	133	88	15	155	32.4	0.262	37	negative
7	161	86	0	0	30.4	0.165	47	positive
2	108	80	0	0	27	0.259	52	positive
7	136	74	26	135	26	0.647	51	negative
5	155	84	44	545	38.7	0.619	34	negative
1	119	86	39	220	45.6	0.808	29	positive
4	96	56	17	49	20.8	0.34	26	negative
5	108	72	43	75	36.1	0.263	33	negative
0	78	88	29	40	36.9	0.434	21	negative
0	107	62	30	74	36.6	0.757	25	positive
2	128	78	37	182	43.3	1.224	31	positive
1	128	48	45	194	40.5	0.613	24	positive
0	161	50	0	0	21.9	0.254	65	negative
6	151	62	31	120	35.5	0.692	28	negative
2	146	70	38	360	28	0.337	29	positive
0	126	84	29	215	30.7	0.52	24	negative
14	100	78	25	184	36.6	0.412	46	positive
8	112	72	0	0	23.6	0.84	58	negative

0	167	0	0	0	32.3	0.839	30	positive
2	144	58	33	135	31.6	0.422	25	positive
5	77	82	41	42	35.8	0.156	35	negative
5	115	98	0	0	52.9	0.209	28	positive
3	150	76	0	0	21	0.207	37	negative
2	120	76	37	105	39.7	0.215	29	negative
10	161	68	23	132	25.5	0.326	47	positive
0	137	68	14	148	24.8	0.143	21	negative
0	128	68	19	180	30.5	1.391	25	positive
2	124	68	28	205	32.9	0.875	30	positive
6	80	66	30	0	26.2	0.313	41	negative
0	106	70	37	148	39.4	0.605	22	negative
2	155	74	17	96	26.6	0.433	27	positive
3	113	50	10	85	29.5	0.626	25	negative
7	109	80	31	0	35.9	1.127	43	positive
2	112	68	22	94	34.1	0.315	26	negative
3	99	80	11	64	19.3	0.284	30	negative
3	182	74	0	0	30.5	0.345	29	positive
3	115	66	39	140	38.1	0.15	28	negative
6	194	78	0	0	23.5	0.129	59	positive
4	129	60	12	231	27.5	0.527	31	negative
3	112	74	30	0	31.6	0.197	25	positive
0	124	70	20	0	27.4	0.254	36	positive
13	152	90	33	29	26.8	0.731	43	positive
2	112	75	32	0	35.7	0.148	21	negative
1	157	72	21	168	25.6	0.123	24	negative
1	122	64	32	156	35.1	0.692	30	positive
10	179	70	0	0	35.1	0.2	37	negative
2	102	86	36	120	45.5	0.127	23	positive
6	105	70	32	68	30.8	0.122	37	negative
8	118	72	19	0	23.1	1.476	46	negative
2	87	58	16	52	32.7	0.166	25	negative
1	180	0	0	0	43.3	0.282	41	positive
12	106	80	0	0	23.6	0.137	44	negative

1	95	60	18	58	23.9	0.26	22	negative
0	165	76	43	255	47.9	0.259	26	negative
0	117	0	0	0	33.8	0.932	44	negative
5	115	76	0	0	31.2	0.343	44	positive
9	152	78	34	171	34.2	0.893	33	positive
7	178	84	0	0	39.9	0.331	41	positive
1	130	70	13	105	25.9	0.472	22	negative
1	95	74	21	73	25.9	0.673	36	negative
1	0	68	35	0	32	0.389	22	negative
5	122	86	0	0	34.7	0.29	33	negative
8	95	72	0	0	36.8	0.485	57	negative
8	126	88	36	108	38.5	0.349	49	negative
1	139	46	19	83	28.7	0.654	22	negative
3	116	0	0	0	23.5	0.187	23	negative
3	99	62	19	74	21.8	0.279	26	negative
5	0	80	32	0	41	0.346	37	positive
4	92	80	0	0	42.2	0.237	29	negative
4	137	84	0	0	31.2	0.252	30	negative
3	61	82	28	0	34.4	0.243	46	negative
1	90	62	12	43	27.2	0.58	24	negative
3	90	78	0	0	42.7	0.559	21	negative
9	165	88	0	0	30.4	0.302	49	positive
1	125	50	40	167	33.3	0.962	28	positive
13	129	0	30	0	39.9	0.569	44	positive
12	88	74	40	54	35.3	0.378	48	negative
1	196	76	36	249	36.5	0.875	29	positive
5	189	64	33	325	31.2	0.583	29	positive
5	158	70	0	0	29.8	0.207	63	negative
5	103	108	37	0	39.2	0.305	65	negative
4	146	78	0	0	38.5	0.52	67	positive
4	147	74	25	293	34.9	0.385	30	negative
5	99	54	28	83	34	0.499	30	negative
6	124	72	0	0	27.6	0.368	29	positive
0	101	64	17	0	21	0.252	21	negative

3	81	86	16	66	27.5	0.306	22	negative
1	133	102	28	140	32.8	0.234	45	positive
3	173	82	48	465	38.4	2.137	25	positive
0	118	64	23	89	0	1.731	21	negative
0	84	64	22	66	35.8	0.545	21	negative
2	105	58	40	94	34.9	0.225	25	negative
2	122	52	43	158	36.2	0.816	28	negative
12	140	82	43	325	39.2	0.528	58	positive
0	98	82	15	84	25.2	0.299	22	negative
1	87	60	37	75	37.2	0.509	22	negative
4	156	75	0	0	48.3	0.238	32	positive
0	93	100	39	72	43.4	1.021	35	negative
1	107	72	30	82	30.8	0.821	24	negative
0	105	68	22	0	20	0.236	22	negative
1	109	60	8	182	25.4	0.947	21	negative
1	90	62	18	59	25.1	1.268	25	negative
1	125	70	24	110	24.3	0.221	25	negative
1	119	54	13	50	22.3	0.205	24	negative
5	116	74	29	0	32.3	0.66	35	positive
8	105	100	36	0	43.3	0.239	45	positive
5	144	82	26	285	32	0.452	58	positive
3	100	68	23	81	31.6	0.949	28	negative
1	100	66	29	196	32	0.444	42	negative
5	166	76	0	0	45.7	0.34	27	positive
1	131	64	14	415	23.7	0.389	21	negative
4	116	72	12	87	22.1	0.463	37	negative
4	158	78	0	0	32.9	0.803	31	positive
2	127	58	24	275	27.7	1.6	25	negative
3	96	56	34	115	24.7	0.944	39	negative
0	131	66	40	0	34.3	0.196	22	positive
3	82	70	0	0	21.1	0.389	25	negative
3	193	70	31	0	34.9	0.241	25	positive
4	95	64	0	0	32	0.161	31	positive
6	137	61	0	0	24.2	0.151	55	negative

5	136	84	41	88	35	0.286	35	positive
9	72	78	25	0	31.6	0.28	38	negative
5	168	64	0	0	32.9	0.135	41	positive
2	123	48	32	165	42.1	0.52	26	negative
4	115	72	0	0	28.9	0.376	46	positive
0	101	62	0	0	21.9	0.336	25	negative
8	197	74	0	0	25.9	1.191	39	positive
1	172	68	49	579	42.4	0.702	28	positive
6	102	90	39	0	35.7	0.674	28	negative
1	112	72	30	176	34.4	0.528	25	negative
1	143	84	23	310	42.4	1.076	22	negative
1	143	74	22	61	26.2	0.256	21	negative
0	138	60	35	167	34.6	0.534	21	positive
3	173	84	33	474	35.7	0.258	22	positive
1	97	68	21	0	27.2	1.095	22	negative
4	144	82	32	0	38.5	0.554	37	positive
1	83	68	0	0	18.2	0.624	27	negative
3	129	64	29	115	26.4	0.219	28	positive
1	119	88	41	170	45.3	0.507	26	negative
2	94	68	18	76	26	0.561	21	negative
0	102	64	46	78	40.6	0.496	21	negative
2	115	64	22	0	30.8	0.421	21	negative
8	151	78	32	210	42.9	0.516	36	positive
4	184	78	39	277	37	0.264	31	positive
0	94	0	0	0	0	0.256	25	negative
1	181	64	30	180	34.1	0.328	38	positive
0	135	94	46	145	40.6	0.284	26	negative
1	95	82	25	180	35	0.233	43	positive
2	99	0	0	0	22.2	0.108	23	negative
3	89	74	16	85	30.4	0.551	38	negative
1	80	74	11	60	30	0.527	22	negative
2	139	75	0	0	25.6	0.167	29	negative
1	90	68	8	0	24.5	1.138	36	negative
0	141	0	0	0	42.4	0.205	29	positive

12	140	85	33	0	37.4	0.244	41	negative
5	147	75	0	0	29.9	0.434	28	negative
1	97	70	15	0	18.2	0.147	21	negative
6	107	88	0	0	36.8	0.727	31	negative
0	189	104	25	0	34.3	0.435	41	positive
2	83	66	23	50	32.2	0.497	22	negative
4	117	64	27	120	33.2	0.23	24	negative
8	108	70	0	0	30.5	0.955	33	positive
4	117	62	12	0	29.7	0.38	30	positive
0	180	78	63	14	59.4	2.42	25	positive
1	100	72	12	70	25.3	0.658	28	negative
0	95	80	45	92	36.5	0.33	26	negative
0	104	64	37	64	33.6	0.51	22	positive
0	120	74	18	63	30.5	0.285	26	negative
1	82	64	13	95	21.2	0.415	23	negative
2	134	70	0	0	28.9	0.542	23	positive
0	91	68	32	210	39.9	0.381	25	negative
2	119	0	0	0	19.6	0.832	72	negative
2	100	54	28	105	37.8	0.498	24	negative
14	175	62	30	0	33.6	0.212	38	positive
1	135	54	0	0	26.7	0.687	62	negative
5	86	68	28	71	30.2	0.364	24	negative
10	148	84	48	237	37.6	1.001	51	positive
9	134	74	33	60	25.9	0.46	81	negative
9	120	72	22	56	20.8	0.733	48	negative
1	71	62	0	0	21.8	0.416	26	negative
8	74	70	40	49	35.3	0.705	39	negative
5	88	78	30	0	27.6	0.258	37	negative
10	115	98	0	0	24	1.022	34	negative
0	124	56	13	105	21.8	0.452	21	negative
0	74	52	10	36	27.8	0.269	22	negative
0	97	64	36	100	36.8	0.6	25	negative
8	120	0	0	0	30	0.183	38	positive
6	154	78	41	140	46.1	0.571	27	negative

1	144	82	40	0	41.3	0.607	28	negative
0	137	70	38	0	33.2	0.17	22	negative
0	119	66	27	0	38.8	0.259	22	negative
7	136	90	0	0	29.9	0.21	50	negative
4	114	64	0	0	28.9	0.126	24	negative
0	137	84	27	0	27.3	0.231	59	negative
2	105	80	45	191	33.7	0.711	29	positive
7	114	76	17	110	23.8	0.466	31	negative
8	126	74	38	75	25.9	0.162	39	negative
4	132	86	31	0	28	0.419	63	negative
3	158	70	30	328	35.5	0.344	35	positive
0	123	88	37	0	35.2	0.197	29	negative
4	85	58	22	49	27.8	0.306	28	negative
0	84	82	31	125	38.2	0.233	23	negative
0	145	0	0	0	44.2	0.63	31	positive
0	135	68	42	250	42.3	0.365	24	positive
1	139	62	41	480	40.7	0.536	21	negative
0	173	78	32	265	46.5	1.159	58	negative
4	99	72	17	0	25.6	0.294	28	negative
8	194	80	0	0	26.1	0.551	67	negative
2	83	65	28	66	36.8	0.629	24	negative
2	89	90	30	0	33.5	0.292	42	negative
4	99	68	38	0	32.8	0.145	33	negative
4	125	70	18	122	28.9	1.144	45	positive
3	80	0	0	0	0	0.174	22	negative
6	166	74	0	0	26.6	0.304	66	negative
5	110	68	0	0	26	0.292	30	negative
2	81	72	15	76	30.1	0.547	25	negative
7	195	70	33	145	25.1	0.163	55	positive
6	154	74	32	193	29.3	0.839	39	negative
2	117	90	19	71	25.2	0.313	21	negative
3	84	72	32	0	37.2	0.267	28	negative
6	0	68	41	0	39	0.727	41	positive
7	94	64	25	79	33.3	0.738	41	negative

3	96	78	39	0	37.3	0.238	40	negative
10	75	82	0	0	33.3	0.263	38	negative
0	180	90	26	90	36.5	0.314	35	positive
1	130	60	23	170	28.6	0.692	21	negative
2	84	50	23	76	30.4	0.968	21	negative
8	120	78	0	0	25	0.409	64	negative
12	84	72	31	0	29.7	0.297	46	positive
0	139	62	17	210	22.1	0.207	21	negative

Test

1	199	76	43	0	42.9	1.394	22	positive
0	198	66	32	274	41.3	0.502	28	positive
2	197	70	99	0	34.7	0.575	62	positive
6	195	70	0	0	30.9	0.328	31	positive
6	190	92	0	0	35.5	0.278	66	positive
4	189	110	31	0	28.5	0.68	37	negative
0	188	82	14	185	32	0.682	22	positive
5	187	76	27	207	43.6	1.034	53	positive
7	187	50	33	392	33.9	0.826	34	positive
3	187	70	22	200	36.4	0.408	36	positive
8	186	90	35	225	34.5	0.423	37	positive
6	183	94	0	0	40.8	1.461	45	negative
4	183	0	0	0	28.4	0.212	36	positive
1	181	78	42	293	40	1.258	22	positive
0	181	88	44	510	43.3	0.222	26	positive
0	179	50	36	159	37.8	0.455	22	positive
3	176	86	27	156	33.3	1.154	52	positive
2	175	88	0	0	22.9	0.326	22	negative
3	174	58	22	194	32.9	0.593	36	positive
2	174	88	37	120	44.5	0.646	24	positive
1	173	74	0	0	36.8	0.088	38	positive
3	173	78	39	185	33.8	0.97	31	positive
9	170	74	31	0	44	0.403	43	positive
3	169	74	19	125	29.9	0.268	31	positive

7	168	88	42	321	38.2	0.787	40	positive
1	168	88	29	0	35	0.905	52	positive
1	167	74	17	144	23.4	0.447	33	positive
8	167	106	46	231	37.6	0.165	43	positive
6	165	68	26	168	33.6	0.631	49	negative
1	164	82	43	67	32.8	0.341	50	negative
9	164	78	0	0	32.8	0.148	45	positive
3	163	70	18	105	31.6	0.268	28	positive
10	162	84	0	0	27.7	0.182	54	negative
0	162	76	36	0	49.6	0.364	26	positive
6	162	62	0	0	24.3	0.178	50	positive
3	158	64	13	387	31.2	0.295	24	negative
13	158	114	0	0	42.3	0.257	44	positive
2	157	74	35	440	39.4	0.134	30	negative
9	156	86	0	0	24.8	0.23	53	positive
2	155	52	27	540	38.7	0.24	25	positive
4	154	72	29	126	31.3	0.338	37	negative
9	154	78	30	100	30.9	0.164	45	negative
8	154	78	32	0	32.4	0.443	45	positive
13	153	88	37	140	40.6	1.174	39	negative
0	152	82	39	272	41.5	0.27	27	negative
0	151	90	46	0	42.1	0.371	21	positive
7	150	78	29	126	35.2	0.692	54	positive
1	149	68	29	127	29.3	0.349	42	positive
6	147	80	0	0	29.5	0.178	50	positive
1	147	94	41	0	49.3	0.358	27	positive
9	145	88	34	165	30.3	0.771	53	positive
9	145	80	46	130	37.9	0.637	40	positive
4	145	82	18	0	32.5	0.235	70	positive
1	144	82	46	180	46.1	0.335	46	positive
1	143	86	30	330	30.1	0.892	23	negative
8	143	66	0	0	34.9	0.129	41	positive
7	142	90	24	480	30.4	0.128	43	positive
0	141	84	26	0	32.4	0.433	22	negative

1	140	74	26	180	24.1	0.828	23	negative
9	140	94	0	0	32.7	0.734	45	positive
11	138	74	26	144	36.1	0.557	50	positive
7	137	90	41	0	32	0.391	39	negative
5	136	82	0	0	0	0.64	69	negative
11	136	84	35	130	28.3	0.26	42	positive
4	136	70	0	0	31.2	1.182	22	positive
0	134	58	20	291	26.4	0.352	21	negative
10	133	68	0	0	27	0.245	36	negative
0	132	78	0	0	32.4	0.393	21	negative
4	132	0	0	0	32.9	0.302	23	positive
3	132	80	0	0	34.4	0.402	44	positive
4	131	68	21	166	33.1	0.16	28	negative
2	130	96	0	0	22.6	0.268	21	negative
3	130	64	0	0	23.1	0.314	22	negative
9	130	70	0	0	34.2	0.652	45	positive
3	130	78	23	79	28.4	0.323	34	positive
6	129	90	7	326	19.6	0.582	60	negative
2	129	74	26	205	33.2	0.591	25	negative
2	129	0	0	0	38.5	0.304	41	negative
3	129	92	49	155	36.4	0.968	32	positive
7	129	68	49	125	38.5	0.439	43	positive
10	129	62	36	0	41.2	0.441	38	positive
5	128	80	0	0	34.6	0.144	45	negative
1	128	82	17	183	27.5	0.115	22	negative
4	128	70	0	0	34.3	0.303	24	negative
3	128	72	25	190	32.4	0.549	27	positive
1	128	88	39	110	36.5	1.057	37	positive
0	127	80	37	210	36.3	0.804	23	negative
11	127	106	0	0	39	0.19	51	negative
4	127	88	11	155	34.5	0.598	28	negative
2	127	46	21	335	34.4	0.176	22	negative
5	126	78	27	22	29.6	0.439	40	negative
0	126	86	27	120	27.4	0.515	21	negative

1	126	60	0	0	30.1	0.349	47	positive
7	125	86	0	0	37.6	0.304	51	negative
3	125	58	0	0	31.6	0.151	24	negative
0	125	68	0	0	24.7	0.206	21	negative
6	125	76	0	0	33.8	0.121	54	positive
4	125	80	0	0	32.3	0.536	27	positive
6	125	78	31	0	27.6	0.565	49	positive
3	124	80	33	130	33.2	0.305	26	negative
7	124	70	33	215	25.5	0.161	37	negative
1	124	74	36	0	27.8	0.1	30	negative
1	124	60	32	0	35.8	0.514	21	negative
8	124	76	24	600	28.7	0.687	52	positive
6	123	72	45	230	33.6	0.733	34	negative
5	123	74	40	77	34.1	0.269	28	negative
3	123	100	35	240	57.3	0.88	22	negative
0	123	72	0	0	36.3	0.258	52	positive
2	122	60	18	106	29.8	0.717	22	negative
2	122	76	27	200	35.9	0.483	26	negative
2	122	70	27	0	36.8	0.34	27	negative
12	121	78	17	0	26.5	0.259	62	negative
2	121	70	32	95	39.1	0.886	23	negative
1	121	78	39	74	39	0.261	28	negative
5	121	72	23	112	26.2	0.245	30	negative
0	121	66	30	165	34.3	0.203	33	positive
3	121	52	0	0	36	0.127	25	positive
2	120	54	0	0	26.8	0.455	27	negative
1	120	80	48	200	38.9	1.162	41	negative
8	120	86	0	0	28.4	0.259	22	positive
11	120	80	37	150	42.3	0.785	48	positive
1	119	44	47	63	35.5	0.28	25	negative
0	119	0	0	0	32.4	0.141	24	positive
4	118	70	0	0	44.5	0.904	26	negative
2	118	80	0	0	42.9	0.693	21	positive
0	117	66	31	188	30.8	0.493	22	negative

6	117	96	0	0	28.7	0.157	30	negative
1	117	60	23	106	33.8	0.466	27	negative
5	117	86	30	105	39.1	0.251	42	negative
3	116	74	15	105	26.3	0.107	24	negative
1	116	70	28	0	27.4	0.204	21	negative
1	116	78	29	180	36.1	0.496	25	negative
6	115	60	39	0	33.7	0.245	40	positive
10	115	0	0	0	0	0.261	30	positive
6	114	0	0	0	0	0.189	26	negative
6	114	88	0	0	27.8	0.247	66	negative
1	114	66	36	200	38.1	0.289	21	negative
7	114	64	0	0	27.4	0.732	34	positive
2	112	78	50	140	39.4	0.175	24	negative
2	112	86	42	160	38.4	0.246	28	negative
1	112	80	45	132	34.8	0.217	24	negative
4	112	78	40	0	39.4	0.236	38	negative
9	112	82	24	0	28.2	1.282	50	positive
0	111	65	0	0	24.6	0.66	31	negative
3	111	58	31	44	29.5	0.43	22	negative
1	111	62	13	182	24	0.138	23	negative
2	111	60	0	0	26.2	0.343	23	negative
1	111	94	0	0	32.8	0.265	45	negative
11	111	84	40	0	46.8	0.925	45	positive
10	111	70	27	0	27.5	0.141	40	positive
8	110	76	0	0	27.8	0.237	58	negative
4	110	76	20	100	28.4	0.118	27	negative
6	109	60	27	0	25	0.206	27	negative
1	109	38	18	120	23.1	0.407	26	negative
1	109	58	18	116	28.5	0.219	22	negative
6	108	44	20	130	24	0.813	35	negative
1	108	88	19	0	27.1	0.4	24	negative
2	108	64	0	0	30.8	0.158	21	negative
1	108	60	46	178	35.5	0.415	24	negative
3	108	62	24	0	26	0.223	25	negative

0	107	76	0	0	45.3	0.686	24	negative
0	107	60	25	0	26.4	0.133	23	negative
1	107	50	19	0	28.3	0.181	29	negative
8	107	80	0	0	24.6	0.856	34	negative
3	106	54	21	158	30.9	0.292	24	negative
3	106	72	0	0	25.8	0.207	27	negative
1	106	70	28	135	34.2	0.142	22	negative
2	106	56	27	165	29	0.426	22	negative
1	106	76	0	0	37.5	0.197	26	negative
0	105	90	0	0	29.6	0.197	46	negative
6	105	80	28	0	32.5	0.878	26	negative
2	105	75	0	0	23.3	0.56	53	negative
5	104	74	0	0	28.8	0.153	48	negative
13	104	72	0	0	31.2	0.465	38	positive
11	103	68	40	0	46.2	0.126	42	negative
6	103	66	0	0	24.3	0.249	29	negative
3	103	72	30	152	27.6	0.73	27	negative
0	102	78	40	90	34.5	0.238	24	negative
0	102	86	17	105	29.3	0.695	27	negative
3	102	74	0	0	29.5	0.121	32	negative
3	102	44	20	94	30.8	0.4	26	negative
1	102	74	0	0	39.5	0.293	42	positive
2	101	58	35	90	21.8	0.155	22	negative
2	101	58	17	265	24.2	0.614	23	negative
10	101	76	48	180	32.9	0.171	63	negative
8	100	76	0	0	38.7	0.19	42	negative
1	100	74	12	46	19.5	0.149	28	negative
12	100	84	33	105	30	0.488	46	negative
8	100	74	40	215	39.4	0.661	43	positive
3	99	54	19	86	25.6	0.154	24	negative
6	99	60	19	54	26.9	0.497	32	negative
1	99	72	30	18	38.6	0.412	21	negative
1	99	58	10	0	25.4	0.551	21	negative
0	99	0	0	0	25	0.253	22	negative

2	99	60	17	160	36.6	0.453	21	negative
2	98	60	17	120	34.7	0.198	22	negative
6	98	58	33	190	34	0.43	43	negative
1	97	64	19	82	18.2	0.299	21	negative
1	97	70	40	0	38.1	0.218	30	negative
7	97	76	32	91	40.9	0.871	32	positive
5	97	76	27	0	35.6	0.378	52	positive
6	96	0	0	0	23.7	0.19	28	negative
2	95	54	14	88	26.1	0.748	22	negative
0	95	64	39	105	44.6	0.366	22	negative
4	95	60	32	0	35.4	0.284	28	negative
0	94	70	27	115	43.5	0.347	21	negative
4	94	65	22	0	24.7	0.148	21	negative
2	94	76	18	66	31.6	0.649	23	negative
10	94	72	18	0	23.1	0.595	56	negative
1	93	56	11	0	22.5	0.417	22	negative
0	93	60	0	0	35.3	0.263	25	negative
1	93	70	31	0	30.4	0.315	23	negative
2	93	64	32	160	38	0.674	23	positive
6	92	62	32	126	32	0.085	46	negative
1	92	62	25	41	19.5	0.482	25	negative
2	92	76	20	0	24.2	1.698	28	negative
10	92	62	0	0	25.9	0.167	31	negative
2	92	52	0	0	30.1	0.141	22	negative
9	91	68	0	0	24.2	0.2	58	negative
2	91	62	0	0	27.3	0.525	22	negative
6	91	0	0	0	29.8	0.501	31	negative
0	91	80	0	0	32.4	0.601	27	negative
1	91	54	25	100	25.2	0.234	23	negative
8	91	82	0	0	35.6	0.587	68	negative
4	90	88	47	54	37.7	0.362	29	negative
4	90	0	0	0	28	0.61	31	negative
2	90	60	0	0	23.5	0.191	25	negative
10	90	85	32	0	34.9	0.825	56	positive

1	89	24	19	25	27.8	0.559	21	negative
9	89	62	0	0	22.5	0.142	33	negative
1	88	78	29	76	32	0.365	29	negative
1	88	62	24	44	29.9	0.422	23	negative
2	88	58	26	16	28.4	0.766	22	negative
3	87	60	18	0	21.8	0.444	21	negative
1	87	68	34	77	37.6	0.401	24	negative
1	86	66	52	65	41.3	0.917	29	negative
11	85	74	0	0	30.1	0.3	35	negative
4	84	90	23	56	39.5	0.159	25	negative
3	84	68	30	106	31.9	0.591	25	negative
1	84	64	23	115	36.9	0.471	28	negative
4	83	86	19	0	29.3	0.317	34	negative
2	82	52	22	115	28.5	1.699	25	negative
1	81	74	41	57	46.3	1.096	32	negative
6	80	80	36	0	39.8	0.177	28	negative
3	80	82	31	70	34.2	1.292	27	positive
3	78	70	0	0	32.5	0.27	39	negative
1	77	56	30	56	33.3	1.251	24	negative
13	76	60	0	0	32.8	0.18	41	negative
0	73	0	0	0	21.1	0.342	25	negative
2	68	70	32	66	25	0.187	25	negative
2	68	62	13	15	20.1	0.257	23	negative
10	68	106	23	49	35.5	0.285	47	negative
0	67	76	0	0	45.3	0.194	46	negative
8	65	72	23	0	32	0.6	42	negative
0	57	60	0	0	21.7	0.735	67	negative
2	56	56	28	45	24.2	0.332	22	negative

Appendix D

Iris (Training and Test)

Training

SepalLength	SepalWidth	PetalLength	PetalWidth	Class
4.7	3.2	1.6	0.2	Iris-setosa
7.2	3	5.8	1.6	Iris-virginica
6.7	3.1	4.4	1.4	Iris-versicolor
5.1	3.3	1.7	0.5	Iris-setosa
5.7	2.8	4.1	1.3	Iris-versicolor
5.4	3.9	1.3	0.4	Iris-setosa
7.4	2.8	6.1	1.9	Iris-virginica
6.5	3	5.8	2.2	Iris-virginica
6.6	3	4.4	1.4	Iris-versicolor
5.7	2.9	4.2	1.3	Iris-versicolor
5.6	2.5	3.9	1.1	Iris-versicolor
5.1	2.5	3	1.1	Iris-versicolor
5	3.4	1.5	0.2	Iris-setosa
6	2.2	4	1	Iris-versicolor
7.7	2.6	6.9	2.3	Iris-virginica
5	3.5	1.3	0.3	Iris-setosa
5.7	2.5	5	2	Iris-virginica
5.1	3.8	1.5	0.3	Iris-setosa
7	3.2	4.7	1.4	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
6	2.2	5	1.5	Iris-virginica
5.6	3	4.5	1.5	Iris-versicolor
7.3	2.9	6.3	1.8	Iris-virginica
5.7	4.4	1.5	0.4	Iris-setosa
5.1	3.8	1.9	0.4	Iris-setosa
5.7	2.6	3.5	1	Iris-versicolor
5.8	2.7	4.1	1	Iris-versicolor
6	3	4.8	1.8	Iris-virginica

5.8	2.7	3.9	1.2	Iris-versicolor
6.5	3	5.5	1.8	Iris-virginica
6.7	3	5.2	2.3	Iris-virginica
5.2	4.1	1.5	0.1	Iris-setosa
6.4	2.9	4.3	1.3	Iris-versicolor
6.7	2.5	5.8	1.8	Iris-virginica
6.4	3.1	5.5	1.8	Iris-virginica
6.3	3.3	6	2.5	Iris-virginica
5.9	3	5.1	1.8	Iris-virginica
5	3.3	1.4	0.2	Iris-setosa
5.6	2.8	4.9	2	Iris-virginica
6.8	3.2	5.9	2.3	Iris-virginica
4.6	3.1	1.5	0.2	Iris-setosa
6.3	2.5	5	1.9	Iris-virginica
6.5	3.2	5.1	2	Iris-virginica
5	2.3	3.3	1	Iris-versicolor
5.1	3.4	1.5	0.2	Iris-setosa
4.9	2.4	3.3	1	Iris-versicolor
4.8	3.1	1.6	0.2	Iris-setosa
7.2	3.6	6.1	2.5	Iris-virginica
5.1	3.8	1.6	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.6	3.2	1.4	0.2	Iris-setosa
6.9	3.1	5.4	2.1	Iris-virginica
5.5	3.5	1.3	0.2	Iris-setosa
5.7	3	4.2	1.2	Iris-versicolor
5	2	3.5	1	Iris-versicolor
5.1	3.7	1.5	0.4	Iris-setosa
5.1	3.5	1.4	0.2	Iris-setosa
6.7	3.1	5.6	2.4	Iris-virginica
6.4	2.8	5.6	2.2	Iris-virginica
7.9	3.8	6.4	2	Iris-virginica
6.7	3.3	5.7	2.1	Iris-virginica
6.1	2.8	4	1.3	Iris-versicolor

5.8	2.7	5.1	1.9	Iris-virginica
5.4	3	4.5	1.5	Iris-versicolor
6.1	2.8	4.7	1.2	Iris-versicolor
4.9	3.1	1.5	0.1	Iris-setosa
6.3	3.4	5.6	2.4	Iris-virginica
5.9	3.2	4.8	1.8	Iris-versicolor
6.4	3.2	5.3	2.3	Iris-virginica
5.4	3.4	1.5	0.4	Iris-setosa
6.3	2.7	4.9	1.8	Iris-virginica
5.5	2.4	3.7	1	Iris-versicolor
6.1	3	4.6	1.4	Iris-versicolor
5.4	3.9	1.7	0.4	Iris-setosa
6.9	3.1	5.1	2.3	Iris-virginica
7.7	3.8	6.7	2.2	Iris-virginica
5.2	3.4	1.4	0.2	Iris-setosa
6.3	3.3	4.7	1.6	Iris-versicolor
5.7	2.8	4.5	1.3	Iris-versicolor
6.1	2.6	5.6	1.4	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica
5.8	4	1.2	0.2	Iris-setosa
6.6	2.9	4.6	1.3	Iris-versicolor
4.9	2.5	4.5	1.7	Iris-virginica
5.2	3.5	1.5	0.2	Iris-setosa
5.8	2.6	4	1.2	Iris-versicolor
4.3	3	1.1	0.1	Iris-setosa
7.1	3	5.9	2.1	Iris-virginica
5.5	2.3	4	1.3	Iris-versicolor
4.6	3.6	1	0.2	Iris-setosa
5.5	4.2	1.4	0.2	Iris-setosa
5.2	2.7	3.9	1.4	Iris-versicolor
4.9	3.1	1.5	0.1	Iris-setosa
7.7	3	6.1	2.3	Iris-virginica
6.2	2.8	4.8	1.8	Iris-virginica

5.5	2.6	4.4	1.2	Iris-versicolor
5.4	3.4	1.7	0.2	Iris-setosa
6	3.4	4.5	1.6	Iris-versicolor
6.5	3	5.2	2	Iris-virginica
6.8	2.8	4.8	1.4	Iris-versicolor

Test

6.7	3	5	1.7	Iris-versicolor
6.4	2.7	5.3	1.9	Iris-virginica
7.2	3.2	6	1.8	Iris-virginica
6	2.9	4.5	1.5	Iris-versicolor
4.8	3.4	1.9	0.2	Iris-setosa
4.4	3.2	1.3	0.2	Iris-setosa
7.7	2.8	6.7	2	Iris-virginica
5.4	3.7	1.5	0.2	Iris-setosa
5.6	2.9	3.6	1.3	Iris-versicolor
5	3.4	1.6	0.4	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.5	2.3	1.3	0.3	Iris-setosa
6.5	2.8	4.6	1.5	Iris-versicolor
6.3	2.5	4.9	1.5	Iris-versicolor
6.8	3	5.5	2.1	Iris-virginica
5	3	1.6	0.2	Iris-setosa
6.4	2.8	5.6	2.1	Iris-virginica
4.4	3	1.3	0.2	Iris-setosa
6.2	3.4	5.4	2.3	Iris-virginica
5	3.5	1.6	0.6	Iris-setosa
7.6	3	6.6	2.1	Iris-virginica
5.6	3	4.1	1.3	Iris-versicolor
4.8	3	1.4	0.3	Iris-setosa
6.1	2.9	4.7	1.4	Iris-versicolor
6.2	2.2	4.5	1.5	Iris-versicolor

6.7	3.1	4.7	1.5	Iris-versicolor
5.5	2.5	4	1.3	Iris-versicolor
6.9	3.2	5.7	2.3	Iris-virginica
5	3.6	1.4	0.2	Iris-setosa
6.3	2.9	5.6	1.8	Iris-virginica
4.9	3.1	1.5	0.1	Iris-setosa
6.4	3.2	4.5	1.5	Iris-versicolor
5.5	2.4	3.8	1.1	Iris-versicolor
5	3.2	1.2	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
6.1	3	4.9	1.8	Iris-virginica
5.7	3.8	1.7	0.3	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
5.8	2.8	5.1	2.4	Iris-virginica
5.6	2.7	4.2	1.3	Iris-versicolor
6.2	2.9	4.3	1.3	Iris-versicolor
6	2.7	5.1	1.6	Iris-versicolor
4.8	3	1.4	0.1	Iris-setosa
6.7	3.3	5.7	2.5	Iris-virginica
6.3	2.8	5.1	1.5	Iris-virginica
5.9	3	4.2	1.5	Iris-versicolor
5.1	3.5	1.4	0.3	Iris-setosa
6.3	2.3	4.4	1.3	Iris-versicolor
4.6	3.4	1.4	0.3	Iris-setosa

Appendix E

Bumpus Sparrow (Training and Test)

Training

ID	total_length	Alar_length	Length_bead_head	Length_humerus	Length_keel_sternum	S/D
10	158	238	31	18.8	22	S
48	162	245	32.5	18.5	21.1	D
11	158	240	31.3	18.6	22	S
47	153	237	30.6	18.6	20.4	D
40	163	249	33.4	19.5	22.8	D
18	153	238	30.5	18.2	20.9	S
24	160	242	32.6	18.8	21.7	D
44	161	245	32.1	19.1	20.8	D
45	155	235	30.7	17.7	19.6	D
2	154	240	30.4	17.9	19.6	S
4	153	236	30.9	17.7	20.2	S
14	157	245	32	19.1	20	S
22	155	240	31.4	18	20.7	D
26	160	250	31.7	18.8	22.5	D
34	159	247	30.9	18.1	19	D
35	155	243	30.9	18.5	21.3	D
13	161	246	32.3	19.3	21.8	S
32	162	243	31.6	18.8	21.3	D
9	164	248	32.7	19.1	21.1	S
21	159	236	31.5	18	21.5	S
36	162	252	31.9	19.1	22.2	D

42	156	237	31.7	18.2	20.3	D
28	157	245	32.2	19.5	21.4	D
7	157	238	30.9	18.4	20.2	S
16	156	237	30.9	18	20.3	S
12	160	244	31.1	18.6	20.5	S
15	157	235	31.5	18.1	19.8	S
43	159	238	31.5	18.4	20.3	D
46	162	247	31.9	19.1	20.4	D
29	165	245	33.1	19.8	22.7	D
27	155	237	31	18.5	20	D
25	152	232	30.3	17.2	19.8	D
49	164	248	32.3	18.8	20.9	D
Test						
38	159	242	30.8	18.2	20.5	D
3	153	240	31	18.4	20.6	S
33	159	245	31.8	18.5	21.7	D
39	155	238	31.2	17.9	19.3	D
30	153	231	30.1	17.3	19.8	D
23	156	240	31.5	18.2	20.6	D
1	156	245	31.6	18.5	20.5	S
19	155	236	30.3	18.5	20.1	S
41	163	242	31	18.1	20.7	D
37	152	230	30.4	17.3	18.6	D
6	163	247	32	19	20.9	S
17	158	244	31.4	18.5	21.6	S

20	163	246	32.5	18.6	21.9	S
31	162	239	30.3	18	23.1	D
5	155	243	31.5	18.6	20.3	S
8	155	239	32.8	18.6	21.2	S

Appendix F
ILPD (Training and Test)

Training

Age	TB	DB	Alkphos		Sgpt	Sgot	TP	ALB	AG	Class
70	2.7	1.2	365	62	55	6	2.4	0.6	LP	
35	26.3	12.1	108	168	630	9.2	2	0.3	LP	
40	3.9	1.7	350	950	1500	6.7	3.8	1.3	LP	
32	25	13.7	560	41	88	7.9	2.5	2.5	LP	
37	0.8	0.2	205	31	36	9.2	4.6	1	NLP	
33	2.1	0.7	205	50	38	6.8	3	0.7	LP	
10	0.8	0.1	395	25	75	7.6	3.6	0.9	LP	
38	1.7	1	180	18	34	7.2	3.6	1	LP	
32	23	11.3	300	482	275	7.1	3.5	0.9	LP	
66	15.2	7.7	356	321	562	6.5	2.2	0.4	LP	
74	1.1	0.4	214	22	30	8.1	4.1	1	LP	
49	2	0.6	209	48	32	5.7	3	1.1	NLP	
48	3.2	1.6	257	33	116	5.7	2.2	0.62	LP	
24	1	0.2	189	52	31	8	4.8	1.5	LP	
53	0.8	0.2	193	96	57	6.7	3.6	1.16	LP	
55	0.8	0.2	155	21	17	6.9	3.8	1.4	LP	
48	0.7	0.2	208	15	30	4.6	2.1	0.8	NLP	
35	1	0.3	805	133	103	7.9	3.3	0.7	LP	
38	0.7	0.1	152	90	21	7.1	4.2	1.4	NLP	
49	3.9	2.1	189	65	181	6.9	3	0.7	LP	
42	0.7	0.2	152	35	81	6.2	3.2	1.06	LP	
60	0.5	0.1	500	20	34	5.9	1.6	0.37	NLP	
7	27.2	11.8	1420	790	1050	6.1	2	0.4	LP	
47	3	1.5	292	64	67	5.6	1.8	0.47	LP	
42	1	0.3	154	38	21	6.8	3.9	1.3	NLP	
60	2	1.1	664	52	104	6	2.1	0.53	LP	
78	1	0.3	152	28	70	6.3	3.1	0.9	LP	
42	8.9	4.5	272	31	61	5.8	2	0.5	LP	

75	0.9	0.2	206	44	33	6.2	2.9	0.8	LP
48	4.5	2.3	282	13	74	7	2.4	0.52	LP
48	1.4	0.8	621	110	176	7.2	3.9	1.1	LP
51	4	2.5	275	382	330	7.5	4	1.1	LP
27	1	0.3	180	56	111	6.8	3.9	1.85	NLP
85	1	0.3	208	17	15	7	3.6	1	NLP
32	15.9	7	280	1350	1600	5.6	2.8	1	LP
61	0.8	0.1	282	85	231	8.5	4.3	1	LP
42	6.8	3.2	630	25	47	6.1	2.3	0.6	NLP
52	0.9	0.2	156	35	44	4.9	2.9	1.4	LP
58	0.8	0.2	123	56	48	6	3	1	LP
30	0.8	0.2	198	30	58	5.2	2.8	1.1	LP
60	1.4	0.7	159	10	12	4.9	2.5	1	NLP
38	0.7	0.2	110	22	18	6.4	2.5	0.64	LP
40	0.6	0.1	98	35	31	6	3.2	1.1	LP
48	1.6	1	588	74	113	7.3	2.4	0.4	LP
27	1.2	0.4	179	63	39	6.1	3.3	1.1	NLP
58	1	0.5	158	37	43	7.2	3.6	1	LP
60	19.6	9.5	466	46	52	6.1	2	0.4	LP
13	1.5	0.5	575	29	24	7.9	3.9	0.9	LP
75	0.8	0.2	188	20	29	4.4	1.8	0.6	LP
51	0.8	0.2	230	24	46	6.5	3.1	0.9	LP
63	0.5	0.1	170	21	28	5.5	2.5	0.8	LP
64	0.9	0.3	310	61	58	7	3.4	0.9	NLP
53	19.8	10.4	238	39	221	8.1	2.5	0.4	LP
66	16.6	7.6	315	233	384	6.9	2	0.4	LP
29	1.2	0.4	160	20	22	6.2	3	0.9	NLP
60	5.7	2.8	214	412	850	7.3	3.2	0.78	LP
27	1	0.2	205	137	145	6	3	1	LP
72	1.7	0.8	200	28	37	6.2	3	0.93	LP
49	0.6	0.1	218	50	53	5	2.4	0.9	LP
26	1.7	0.6	210	62	56	5.4	2.2	0.6	LP
62	5	2.1	103	18	40	5	2.1	1.72	LP
46	20	10	254	140	540	5.4	3	1.2	LP

32	12.1	6	515	48	92	6.6	2.4	0.5	LP
17	0.9	0.2	279	40	46	7.3	4	1.2	NLP
56	1.1	0.5	180	30	42	6.9	3.8	1.2	NLP
32	3.7	1.6	612	50	88	6.2	1.9	0.4	LP
28	0.6	0.2	159	15	16	7	3.5	1	NLP
28	0.9	0.2	215	50	28	8	4	1	LP
46	3.3	1.5	172	25	41	5.6	2.4	0.7	LP
18	1.3	0.7	316	10	21	6	2.1	0.5	NLP
57	1.4	0.7	470	62	88	5.6	2.5	0.8	LP
60	3.2	1.8	750	79	145	7.8	3.2	0.69	LP
40	1.1	0.3	230	1630	960	4.9	2.8	1.3	LP
58	2.8	1.3	670	48	79	4.7	1.6	0.5	LP
66	0.7	0.2	162	24	20	6.4	3.2	1	NLP
48	0.9	0.2	173	26	27	6.2	3.1	1	LP
45	0.6	0.1	196	29	30	5.8	2.9	1	LP
60	2.2	1	271	45	52	6.1	2.9	0.9	NLP
26	0.6	0.2	120	45	51	7.9	4	1	LP
61	0.7	0.2	145	53	41	5.8	2.7	0.87	LP
21	0.7	0.2	211	14	23	7.3	4.1	1.2	NLP
55	1.8	9	272	22	79	6.1	2.7	0.7	LP
61	0.8	0.2	163	18	19	6.3	2.8	0.8	NLP
12	1	0.2	719	157	108	7.2	3.7	1	LP
70	1.3	0.4	358	19	14	6.1	2.8	0.8	LP
32	0.7	0.2	276	102	190	6	2.9	0.93	LP
34	6.2	3	240	1680	850	7.2	4	1.2	LP
29	0.7	0.2	165	55	87	7.5	4.6	1.58	LP
45	1.7	0.8	315	12	38	6.3	2.1	0.5	LP
55	10.9	5.1	1350	48	57	6.4	2.3	0.5	LP
65	0.7	0.2	265	30	28	5.2	1.8	0.52	NLP
24	0.7	0.2	218	47	26	6.6	3.3	1	LP
60	8.9	4	950	33	32	6.8	3.1	0.8	LP
74	0.6	0.1	272	24	98	5	2	0.6	LP
26	0.9	0.2	154	16	12	7	3.5	1	LP
28	0.6	0.1	137	22	16	4.9	1.9	0.6	NLP

52	0.8	0.2	245	48	49	6.4	3.2	1	LP
57	0.6	0.1	210	51	59	5.9	2.7	0.8	LP
66	11.3	5.6	1110	1250	4929	7	2.4	0.5	LP
54	0.9	0.2	290	15	18	6.1	2.8	0.8	LP
39	1.6	0.8	230	88	74	8	4	1	NLP
54	23.2	12.6	574	43	47	7.2	3.5	0.9	LP
24	3.3	1.6	174	11	33	7.6	3.9	1	NLP
54	2.2	1.2	195	55	95	6	3.7	1.6	LP
48	0.8	0.2	218	32	28	5.2	2.5	0.9	NLP
55	18.4	8.8	206	64	178	6.2	1.8	0.4	LP
55	0.7	0.2	290	53	58	6.8	3.4	1	LP
31	0.9	0.2	518	189	17	5.3	2.3	0.7	LP
27	1.3	0.6	106	25	54	8.5	4.8	0.9	NLP
34	8.7	4	298	58	138	5.8	2.4	0.7	LP
42	30.5	14.2	285	65	130	5.2	2.1	0.6	LP
57	4	1.9	190	45	111	5.2	1.5	0.4	LP
14	0.9	0.3	310	21	16	8.1	4.2	1	NLP
18	1.8	0.7	178	35	36	6.8	3.6	1.1	LP
50	0.7	0.2	192	18	15	7.4	4.2	1.3	NLP
50	1.1	0.3	175	20	19	7.1	4.5	1.7	NLP
41	1.2	0.5	246	34	42	6.9	3.4	0.97	LP
48	0.7	0.2	326	29	17	8.7	5.5	1.7	LP
60	0.6	0.1	186	20	21	6.2	3.3	1.1	NLP
38	0.8	0.2	185	25	21	7	3	0.7	LP
17	0.9	0.2	224	36	45	6.9	4.2	1.55	LP
37	0.7	0.2	176	28	34	5.6	2.6	0.8	LP
58	0.8	0.2	298	33	59	6.2	3.1	1	LP
55	3.6	1.6	349	40	70	7.2	2.9	0.6	LP
16	2.6	1.2	236	131	90	5.4	2.6	0.9	LP
22	0.8	0.2	198	20	26	6.8	3.9	1.3	LP
40	0.9	0.2	285	32	27	7.7	3.5	0.8	LP
14	1.4	0.5	269	58	45	6.7	3.9	1.4	LP
55	8.2	3.9	1350	52	65	6.7	2.9	0.7	LP
51	0.8	0.2	175	48	22	8.1	4.6	1.3	LP

43	1.3	0.6	155	15	20	8	4	1	NLP
65	0.9	0.2	170	33	66	7	3	0.75	LP
50	2.7	1.6	157	149	156	7.9	3.1	0.6	LP
50	0.7	0.2	206	18	17	8.4	4.2	1	NLP
38	2.6	1.2	410	59	57	5.6	3	0.8	NLP
47	0.8	0.2	236	10	13	6.7	2.9	0.76	NLP
30	0.8	0.2	174	21	47	4.6	2.3	1	LP
63	0.9	0.2	194	52	45	6	3.9	1.85	NLP
72	2.7	1.3	260	31	56	7.4	3	0.6	LP
22	6.7	3.2	850	154	248	6.2	2.8	0.8	LP
40	0.9	0.3	293	232	245	6.8	3.1	0.8	LP
51	0.7	0.1	180	25	27	6.1	3.1	1	LP
53	0.9	0.2	210	35	32	8	3.9	0.9	NLP
60	2.9	1.3	230	32	44	5.6	2	0.5	LP
69	0.8	0.2	146	42	70	8.4	4.9	1.4	NLP
8	0.9	0.2	401	25	58	7.5	3.4	0.8	LP
68	0.7	0.2	186	18	15	6.4	3.8	1.4	LP
35	0.7	0.2	198	42	30	6.8	3.4	1	LP
40	30.8	18.3	285	110	186	7.9	2.7	0.5	LP
68	1.8	0.5	151	18	22	6.5	4	1.6	LP
16	7.7	4.1	268	213	168	7.1	4	1.2	LP
47	3.5	1.6	206	32	31	6.8	3.4	1	LP
34	4.1	2	289	875	731	5	2.7	1.1	LP
54	1.4	0.7	195	36	16	7.9	3.7	0.9	NLP
47	0.9	0.2	265	40	28	8	4	1	LP
37	0.7	0.2	235	96	54	9.5	4.9	1	LP
27	0.6	0.2	161	27	28	3.7	1.6	0.76	NLP
48	1	1.4	144	18	14	8.3	4.2	1	LP
75	2.5	1.2	375	85	68	6.4	2.9	0.8	LP
29	0.9	0.3	202	14	11	6.7	3.6	1.1	LP
32	0.6	0.1	237	45	31	7.5	4.3	1.34	LP
55	0.8	0.2	290	139	87	7	3	0.7	LP
44	0.8	0.2	335	148	86	5.6	3	1.1	LP
52	2.7	1.4	251	20	40	6	1.7	0.39	LP

34	0.6	0.1	161	15	19	6.6	3.4	1	LP
35	0.8	0.2	279	20	25	7.2	3.2	0.8	LP
45	0.9	0.3	189	23	33	6.6	3.9	0.9	LP
33	0.8	0.2	135	30	29	7.2	4.4	1.5	NLP
66	2.9	1.3	168	21	38	5.5	1.8	0.4	LP
18	1.4	0.6	215	440	850	5	1.9	0.6	LP
22	1.1	0.3	138	14	21	7	3.8	1.1	NLP
50	1.6	0.8	218	18	20	5.9	2.9	0.96	LP
52	0.6	0.1	178	26	27	6.5	3.6	1.2	NLP
33	2	1	258	194	152	5.4	3	1.25	LP
45	1.1	0.4	92	91	188	7.2	3.8	1.11	LP
45	0.7	0.2	180	18	58	6.7	3.7	1.2	NLP
31	0.6	0.1	175	48	34	6	3.7	1.6	LP
21	1	0.3	142	27	21	6.4	3.5	1.2	NLP
60	0.7	0.2	171	31	26	7	3.5	1	NLP
64	3	1.4	248	46	40	6.5	3.2	0.9	LP
46	0.8	0.2	185	24	15	7.9	3.7	0.8	LP
35	0.6	0.2	180	12	15	5.2	2.7	0.9	NLP
51	2.9	1.2	189	80	125	6.2	3.1	1	LP
57	1.3	0.4	259	40	86	6.5	2.5	0.6	LP
42	0.8	0.2	158	27	23	6.7	3.1	0.8	NLP
18	0.6	0.2	538	33	34	7.5	3.2	0.7	LP
26	0.7	0.2	144	36	33	8.2	4.3	1.1	LP
35	0.9	0.2	190	40	35	7.3	4.7	1.8	NLP
75	14.8	9	1020	71	42	5.3	2.2	0.7	LP
18	0.6	0.1	265	97	161	5.9	3.1	1.1	LP
39	1.9	0.9	180	42	62	7.4	4.3	1.38	LP
61	0.8	0.2	192	28	35	6.9	3.4	0.9	NLP
27	0.7	0.2	243	21	23	5.3	2.3	0.7	NLP
51	2.2	1	610	17	28	7.3	2.6	0.55	LP
39	1.9	0.9	180	42	62	7.4	4.3	1.38	LP
58	2.4	1.1	915	60	142	4.7	1.8	0.6	LP
43	22.5	11.8	143	22	143	6.6	2.1	0.46	LP
31	0.8	0.2	198	43	31	7.3	4	1.2	LP

28	0.9	0.2	316	25	23	8.5	5.5	1.8	LP
33	0.7	0.2	256	21	30	8.5	3.9	0.8	LP
45	0.8	0.2	140	24	20	6.3	3.2	1	NLP
53	0.9	0.4	238	17	14	6.6	2.9	0.8	LP
33	7.1	3.7	196	622	497	6.9	3.6	1.09	LP
45	2.4	1.1	168	33	50	5.1	2.6	1	LP
48	0.7	0.2	165	32	30	8	4	1	NLP
60	4	1.9	238	119	350	7.1	3.3	0.8	LP
73	1.8	0.9	220	20	43	6.5	3	0.8	LP
21	0.8	0.2	183	33	57	6.8	3.5	1	NLP
25	0.7	0.1	140	32	25	7.6	4.3	1.3	NLP
60	1.8	0.5	201	45	25	3.9	1.7	0.7	NLP
38	0.8	0.2	247	55	92	7.4	4.3	1.38	NLP
84	0.7	0.2	188	13	21	6	3.2	1.1	NLP
40	1.2	0.6	204	23	27	7.6	4	1.1	LP
40	2.1	1	768	74	141	7.8	4.9	1.6	LP
15	0.8	0.2	380	25	66	6.1	3.7	1.5	LP
29	1	0.3	75	25	26	5.1	2.9	1.3	LP
40	0.9	0.3	196	69	48	6.8	3.1	0.8	LP
52	1.8	0.8	97	85	78	6.4	2.7	0.7	LP
50	0.6	0.2	137	15	16	4.8	2.6	1.1	LP
62	6.8	3	542	116	66	6.4	3.1	0.9	LP
49	1	0.3	230	48	58	8.4	4.2	1	LP
70	1.7	0.5	400	56	44	5.7	3.1	1.1	LP
65	7.9	4.3	282	50	72	6	3	1	LP
36	1.2	0.4	358	160	90	8.3	4.4	1.1	NLP
64	1.4	0.5	298	31	83	7.2	2.6	0.5	LP
21	3.9	1.8	150	36	27	6.8	3.9	1.34	LP
32	32.6	14.1	219	95	235	5.8	3.1	1.1	LP
55	3.3	1.5	214	54	152	5.1	1.8	0.5	LP
60	1.9	0.8	614	42	38	4.5	1.8	0.6	LP
7	0.5	0.1	352	28	51	7.9	4.2	1.1	NLP
62	0.7	0.2	162	12	17	8.2	3.2	0.6	NLP
25	0.7	0.2	185	196	401	6.5	3.9	1.5	LP

36	0.8	0.2	182	31	34	6.4	3.8	1.4	NLP
45	23.3	12.8	1550	425	511	7.7	3.5	0.8	LP
46	1.4	0.4	298	509	623	3.6	1	0.3	LP
29	0.8	0.2	156	12	15	6.8	3.7	1.1	NLP
49	1.3	0.4	206	30	25	6	3.1	1.06	NLP
70	0.7	0.2	237	18	28	5.8	2.5	0.75	NLP
58	0.4	0.1	100	59	126	4.3	2.5	1.4	LP
75	10.6	5	562	37	29	5.1	1.8	0.5	LP
42	11.1	6.1	214	60	186	6.9	2.8	2.8	LP
26	0.6	0.2	142	12	32	5.7	2.4	0.75	LP
75	1.4	0.4	215	50	30	5.9	2.6	0.7	LP
47	0.9	0.2	192	38	24	7.3	4.3	1.4	LP
22	0.8	0.2	300	57	40	7.9	3.8	0.9	NLP
50	5.8	3	661	181	285	5.7	2.3	0.67	NLP
40	0.7	0.2	176	28	43	5.3	2.4	0.8	NLP
42	2.7	1.3	219	60	180	7	3.2	0.8	LP
75	0.9	0.2	162	25	20	6.9	3.7	1.1	LP
39	0.6	0.2	188	28	43	8.1	3.3	0.6	LP
17	0.7	0.2	145	18	36	7.2	3.9	1.18	NLP
32	0.7	0.2	189	22	43	7.4	3.1	0.7	NLP
66	0.8	0.2	165	22	32	4.4	2	0.8	LP
53	0.7	0.1	182	20	33	4.8	1.9	0.6	LP
70	3.1	1.6	198	40	28	5.6	2	0.5	LP
51	0.9	0.2	280	21	30	6.7	3.2	0.8	LP
26	7.1	3.3	258	80	113	6.2	2.9	0.8	LP
12	0.8	0.2	302	47	67	6.7	3.5	1.1	NLP
32	12.7	6.2	194	2000	2946	5.7	3.3	1.3	LP
37	0.8	0.2	147	27	46	5	2.5	1	LP
32	0.6	0.1	176	39	28	6	3	1	LP
68	0.7	0.1	145	20	22	5.8	2.9	1	LP
48	0.7	0.1	1630	74	149	5.3	2	0.6	LP
55	0.6	0.2	220	24	32	5.1	2.4	0.88	LP
34	3.7	2.1	490	115	91	6.5	2.8	0.7	LP
60	6.3	3.2	314	118	114	6.6	3.7	1.27	LP

28	1	0.3	90	18	108	6.8	3.1	0.8	NLP
25	0.8	0.1	130	23	42	8	4	1	LP
45	2.8	1.7	263	57	65	5.1	2.3	0.8	LP
36	0.8	0.2	650	70	138	6.6	3.1	0.8	LP
25	0.6	0.1	183	91	53	5.5	2.3	0.7	NLP
72	0.7	0.2	185	16	22	7.3	3.7	1	NLP
55	0.9	0.2	116	36	16	6.2	3.2	1	NLP
75	2.8	1.3	250	23	29	2.7	0.9	0.5	LP
28	0.8	0.3	190	20	14	4.1	2.4	1.4	LP
41	0.9	0.2	169	22	18	6.1	3	0.9	NLP
45	0.7	0.2	170	21	14	5.7	2.5	0.7	LP
32	22.7	10.2	290	322	113	6.6	2.8	0.7	LP
38	2.2	1	310	119	42	7.9	4.1	1	NLP
38	0.6	0.1	165	22	34	5.9	2.9	0.9	NLP
46	10.2	4.2	232	58	140	7	2.7	0.6	LP
75	0.9	0.2	282	25	23	4.4	2.2	1	LP
60	2.4	1	1124	30	54	5.2	1.9	0.5	LP
55	0.8	0.2	225	14	23	6.1	3.3	1.2	NLP
28	0.8	0.2	309	55	23	6.8	4.1	1.51	LP
49	0.8	0.2	198	23	20	7	4.3	1.5	LP
36	0.8	0.2	158	29	39	6	2.2	0.5	NLP
23	1.1	0.5	191	37	41	7.7	4.3	1.2	NLP
56	0.7	0.1	145	26	23	7	4	1.3	NLP
32	18	8.2	298	1250	1050	5.4	2.6	0.9	LP
66	17.3	8.5	388	173	367	7.8	2.6	0.5	LP
29	0.8	0.2	205	30	23	8.2	4.1	1	LP
45	2.2	1.6	320	37	48	6.8	3.4	1	LP
75	2.9	1.3	218	33	37	3	1.5	1	LP
65	0.7	0.1	187	16	18	6.8	3.3	0.9	LP
60	1.5	0.6	360	230	298	4.5	2	0.8	LP
50	1.2	0.4	282	36	32	7.2	3.9	1.1	LP
22	2.7	1	160	82	127	5.5	3.1	1.2	NLP
26	2	0.9	157	54	68	6.1	2.7	0.8	LP
48	1	0.3	310	37	56	5.9	2.5	0.7	LP

36	1.7	0.5	205	36	34	7.1	3.9	1.2	LP
75	0.8	0.2	205	27	24	4.4	2	0.8	LP
26	2	0.9	195	24	65	7.8	4.3	1.2	LP
22	0.9	0.3	179	18	21	6.7	3.7	1.2	NLP
30	1.6	0.4	332	84	139	5.6	2.7	0.9	LP
46	14.2	7.8	374	38	77	4.3	2	0.8	LP
65	1.9	0.8	170	36	43	3.8	1.4	0.58	NLP
48	1.4	0.6	263	38	66	5.8	2.2	0.61	LP
65	0.8	0.2	201	18	22	5.4	2.9	1.1	NLP
33	2.1	1.3	480	38	22	6.5	3	0.8	LP
70	0.9	0.3	220	53	95	6.1	2.8	0.68	LP
74	0.9	0.3	234	16	19	7.9	4	1	LP
62	1.2	0.4	195	38	54	6.3	3.8	1.5	LP
20	1.1	0.5	128	20	30	3.9	1.9	0.95	NLP
50	0.8	0.2	152	29	30	7.4	4.1	1.3	LP
50	0.9	0.3	901	23	17	6.2	3.5	1.2	LP
42	7.4	3.6	298	52	102	4.6	1.9	0.7	LP
35	2	1.1	226	33	135	6	2.7	0.8	NLP
50	1	0.5	239	16	39	7.5	3.7	0.9	LP
37	0.8	0.2	125	41	39	6.4	3.4	1.1	LP
32	0.7	0.1	240	12	15	7	3	0.7	LP
19	1.4	0.8	178	13	26	8	4.6	1.3	NLP
42	0.8	0.2	168	25	18	6.2	3.1	1	LP
23	1	0.3	212	41	80	6.2	3.1	1	LP
60	11	4.9	750	140	350	5.5	2.1	0.6	LP
38	1	0.3	216	21	24	7.3	4.4	1.5	NLP
35	0.9	0.2	190	25	20	6.4	3.6	1.2	NLP
58	0.8	0.2	130	24	25	7	4	1.3	LP
33	1.8	0.8	196	25	22	8	4	1	LP
45	3.5	1.5	189	63	87	5.6	2.9	1	LP
60	0.8	0.2	286	21	27	7.1	4	1.2	LP
21	18.5	9.5	380	390	500	8.2	4.1	1	LP
24	0.7	0.2	188	11	10	5.5	2.3	0.71	NLP
65	0.8	0.1	146	17	29	5.9	3.2	1.18	NLP

62	0.7	0.2	173	46	47	7.3	4.1	1.2	NLP
40	3.5	1.6	298	68	200	7.1	3.4	0.9	LP
56	17.7	8.8	239	43	185	5.6	2.4	0.7	LP
45	1	0.3	250	48	44	8.6	4.3	1	LP
42	16.4	8.9	245	56	87	5.4	2	0.5	LP
38	1.1	0.3	198	86	150	6.3	3.5	1.2	LP
55	4.4	2.9	230	14	25	7.1	2.1	0.4	LP
60	2.6	1.2	171	42	37	5.4	2.7	1	LP
36	5.3	2.3	145	32	92	5.1	2.6	1	NLP
33	0.9	0.8	680	37	40	5.9	2.6	0.8	LP
30	0.8	0.2	182	46	57	7.8	4.3	1.2	NLP
58	0.7	0.1	172	27	22	6.7	3.2	0.9	LP
4	0.9	0.2	348	30	34	8	4	1	NLP
32	12.7	8.4	190	28	47	5.4	2.6	0.9	LP
42	0.8	0.2	114	21	23	7	3	0.7	NLP
39	3.8	1.5	298	102	630	7.1	3.3	0.8	LP
50	0.9	0.2	202	20	26	7.2	4.5	1.66	LP
16	0.7	0.2	418	28	35	7.2	4.1	1.3	NLP
33	0.8	0.2	198	26	23	8	4	1	NLP
41	2.7	1.3	580	142	68	8	4	1	LP
6	0.6	0.1	289	38	30	4.8	2	0.7	NLP
49	0.8	0.2	158	19	15	6.6	3.6	1.2	NLP
34	4.1	2	289	875	731	5	2.7	1.1	LP
65	1.1	0.3	258	48	40	7	3.9	1.2	NLP
51	0.8	0.2	160	34	20	6.9	3.7	1.1	LP
62	7.3	4.1	490	60	68	7	3.3	0.89	LP
73	1.9	0.7	1750	102	141	5.5	2	0.5	LP
65	0.7	0.2	406	24	45	7.2	3.5	0.9	NLP
69	0.9	0.2	215	32	24	6.9	3	0.7	LP
53	1.6	0.9	178	44	59	6.5	3.9	1.5	NLP
65	1.1	0.5	686	16	46	5.7	1.5	0.35	LP
37	1.8	0.8	215	53	58	6.4	3.8	1.4	LP
61	1.5	0.6	196	61	85	6.7	3.8	1.3	NLP
75	6.7	3.6	458	198	143	6.2	3.2	1	LP

72	0.6	0.1	102	31	35	6.3	3.2	1	LP
46	0.8	0.2	182	20	40	6	2.9	0.9	LP
68	0.6	0.1	1620	95	127	4.6	2.1	0.8	LP
45	0.6	0.2	245	22	24	7.1	3.4	0.9	LP
74	1	0.3	175	30	32	6.4	3.4	1.1	LP
33	1.5	7	505	205	140	7.5	3.9	1	LP
34	0.8	0.2	192	15	12	8.6	4.7	1.2	LP
45	2.3	1.3	282	132	368	7.3	4	1.2	LP
50	1	0.3	191	22	31	7.8	4	1	NLP
60	2.1	1	191	114	247	4	1.6	0.6	LP
62	10.9	5.5	699	64	100	7.5	3.2	0.74	LP
26	1.9	0.8	180	22	19	8.2	4.1	1	NLP
45	1.3	0.6	166	49	42	5.6	2.5	0.8	NLP
22	2.4	1	340	25	21	8.3	4.5	1.1	LP
50	7.3	3.6	1580	88	64	5.6	2.3	0.6	NLP
75	1.8	0.8	405	79	50	6.1	2.9	0.9	LP
48	0.8	0.2	150	25	23	7.5	3.9	1	LP
54	5.5	3.2	350	67	42	7	3.2	0.8	LP
49	0.6	0.1	218	50	53	5	2.4	0.9	LP
26	6.8	3.2	140	37	19	3.6	0.9	0.3	LP
50	2.6	1.2	415	407	576	6.4	3.2	1	LP
13	0.6	0.1	320	28	56	7.2	3.6	1	NLP

Test

11	0.7	0.1	592	26	29	7.1	4.2	1.4	Male	NLP
45	0.6	0.1	270	23	42	5.1	2	0.5	Female	NLP
60	0.7	0.2	174	32	14	7.8	4.2	1.1	Male	NLP
32	0.9	0.3	462	70	82	6.2	3.1	1	Male	LP
24	0.9	0.2	195	40	35	7.4	4.1	1.2	Female	NLP
54	22.6	11.4	558	30	37	7.8	3.4	0.8	Female	LP
60	5.8	2.7	599	43	66	5.4	1.8	0.5	Male	LP

70	1.3	0.3	690	93	40	3.6	2.7	0.7	Male	LP
31	1.1	0.3	190	26	15	7.9	3.8	0.9	Female	LP
65	0.7	0.1	392	20	30	5.3	2.8	1.1	Male	LP
37	1.3	0.4	195	41	38	5.3	2.1	0.6	Male	LP
47	2.7	1.3	275	123	73	6.2	3.3	1.1	Male	LP
60	2.3	0.6	272	79	51	6.6	3.5	1.1	Male	LP
58	1.7	0.8	188	60	84	5.9	3.5	1.4	Male	NLP
45	3.2	1.4	512	50	58	6	2.7	0.8	Male	LP
46	0.7	0.2	224	40	23	7.1	3	0.7	Male	LP
38	1.7	0.7	859	89	48	6	3	1	Male	LP
38	1.8	0.8	342	168	441	7.6	4.4	1.3	Male	LP
36	0.7	0.2	152	21	25	5.9	3.1	1.1	Female	NLP
64	1.1	0.5	145	20	24	5.5	3.2	1.39	Male	NLP
42	0.8	0.2	127	29	30	4.9	2.7	1.2	Male	LP
66	0.6	0.2	100	17	148	5	3.3	1.9	Male	NLP
31	0.8	0.2	215	15	21	7.6	4	1.1	Female	LP
46	0.8	0.2	160	31	40	7.3	3.8	1.1	Male	LP
42	0.8	0.2	198	29	19	6.6	3	0.8	Male	NLP
50	27.7	10.8	380	39	348	7.1	2.3	0.4	Female	LP
58	1.7	0.8	1896	61	83	8	3.9	0.95	Female	LP
50	4.2	2.3	450	69	50	7	3	0.7	Male	LP
45	2.2	0.8	209	25	20	8	4	1	Male	LP
48	5	2.6	555	284	190	6.5	3.3	1	Male	LP
36	0.8	0.2	158	29	39	6	2.2	0.5	Male	NLP
30	0.7	0.2	194	32	36	7.5	3.6	0.92	Female	NLP

55	0.8	0.2	482	112	99	5.7	2.6	0.8	Male	LP
26	42.8	19.7	390	75	138	7.5	2.6	0.5	Male	LP
42	0.8	0.2	195	18	15	6.7	3	0.8	Female	LP
65	4.9	2.7	190	33	71	7.1	2.9	0.7	Male	LP
42	8.9	4.5	272	31	61	5.8	2	0.5	Male	LP
20	0.6	0.2	202	12	13	6.1	3	0.9	Female	NLP
42	0.5	0.1	162	155	108	8.1	4	0.9	Female	LP
60	0.8	0.2	215	24	17	6.3	3	0.9	Male	NLP
48	1.1	0.7	527	178	250	8	4.2	1.1	Female	LP
72	3.9	2	195	27	59	7.3	2.4	0.4	Male	LP
45	2.9	1.4	210	74	68	7.2	3.6	1	Male	LP
60	6.8	3.2	308	404	794	6.8	3	0.7	Male	LP
41	7.5	4.3	149	94	92	6.3	3.1	0.9	Male	LP
38	0.8	0.2	145	19	23	6.1	3.1	1.03	Female	NLP
13	0.7	0.2	350	17	24	7.4	4	1.1	Female	LP
32	0.7	0.2	165	31	29	6.1	3	0.96	Male	NLP
65	1	0.3	202	26	13	5.3	2.6	0.9	Female	NLP
18	0.8	0.2	282	72	140	5.5	2.5	0.8	Male	LP
49	0.7	0.1	148	14	12	5.4	2.8	1	Male	NLP
25	0.9	0.3	159	24	25	6.9	4.4	1.7	Female	NLP
26	0.6	0.1	110	15	20	2.8	1.6	1.3	Male	LP
33	1.6	0.5	165	15	23	7.3	3.5	0.92	Male	NLP
30	1.3	0.4	482	102	80	6.9	3.3	0.9	Male	LP
32	15.6	9.5	134	54	125	5.6	4	2.5	Male	LP
57	0.7	0.2	208	35	97	5.1	2.1	0.7	Male	LP

38	1.5	0.4	298	60	103	6	3	1	Male	NLP
18	0.9	0.3	300	30	48	8	4	1	Male	LP
65	0.7	0.2	199	19	22	6.3	3.6	1.3	Male	NLP
46	1.8	0.7	208	19	14	7.6	4.4	1.3	Male	LP
90	1.1	0.3	215	46	134	6.9	3	0.7	Male	LP
38	0.7	0.2	216	349	105	7	3.5	1	Male	LP
37	0.8	0.2	195	60	40	8.2	5	1.5	Male	NLP
31	1.3	0.5	184	29	32	6.8	3.4	1	Male	LP
36	5.3	2.3	145	32	92	5.1	2.6	1	Male	NLP
64	1.1	0.4	201	18	19	6.9	4.1	1.4	Male	LP
66	4.2	2.1	159	15	30	7.1	2.2	0.4	Female	LP
33	2	1.4	2110	48	89	6.2	3	0.9	Male	LP
21	0.7	0.2	135	27	26	6.4	3.3	1	Male	NLP
40	1.9	1	231	16	55	4.3	1.6	0.6	Male	LP
22	0.6	0.2	202	78	41	8	3.9	0.9	Male	LP
38	3.7	2.2	216	179	232	7.8	4.5	1.3	Male	LP
22	2.2	1	215	159	51	5.5	2.5	0.8	Female	LP
55	14.1	7.6	750	35	63	5	1.6	0.47	Male	LP
75	8	4.6	386	30	25	5.5	1.8	0.48	Male	LP
60	2	0.8	190	45	40	6	2.8	0.8	Male	LP
33	0.7	0.1	168	35	33	7	3.7	1.1	Male	LP
72	0.7	0.1	196	20	35	5.8	2	0.5	Male	LP
46	15.8	7.2	227	67	220	6.9	2.6	0.6	Male	LP
44	0.9	0.2	182	29	82	7.1	3.7	1	Male	NLP
39	6.6	3	215	190	950	4	1.7	0.7	Male	LP

62	0.6	0.1	160	42	110	4.9	2.6	1.1	Male	NLP
65	0.7	0.2	182	23	28	6.8	2.9	0.7	Female	NLP
29	0.7	0.1	162	52	41	5.2	2.5	0.9	Female	NLP
38	2.7	1.4	105	25	21	7.5	4.2	1.2	Male	NLP
26	1	0.3	163	48	71	7.1	3.7	1	Male	NLP
58	1	0.4	182	14	20	6.8	3.4	1	Male	LP
32	30.5	17.1	218	39	79	5.5	2.7	0.9	Male	LP
38	2.6	1.2	410	59	57	5.6	3	0.8	Female	NLP
43	0.9	0.3	140	12	29	7.4	3.5	1.8	Female	LP
55	1.1	0.3	215	21	15	6.2	2.9	0.8	Male	NLP
46	4.7	2.2	310	62	90	6.4	2.5	0.6	Female	LP
57	4.5	2.3	315	120	105	7	4	1.3	Male	LP
35	0.8	0.2	198	36	32	7	4	1.3	Male	NLP
60	11.5	5	1050	99	187	6.2	2.8	0.8	Male	LP
31	0.8	0.2	158	21	16	6	3	1	Female	LP
40	0.7	0.1	202	37	29	5	2.6	1	Male	LP
72	0.8	0.2	148	23	35	6	3	1	Male	LP
48	0.9	0.2	175	24	54	5.5	2.7	0.9	Female	NLP
67	2.2	1.1	198	42	39	7.2	3	0.7	Male	LP
52	0.6	0.1	171	22	16	6.6	3.6	1.2	Male	LP
65	1.4	0.6	260	28	24	5.2	2.2	0.7	Male	NLP
33	3.4	1.6	186	779	844	7.3	3.2	0.7	Male	LP
66	1.1	0.5	167	13	56	7.1	4.1	1.36	Male	LP
42	0.8	0.2	182	22	20	7.2	3.9	1.1	Female	LP
45	0.7	0.2	153	41	42	4.5	2.2	0.9	Female	NLP

60	5.8	3	257	107	104	6.6	3.5	1.12	Male	LP
4	0.8	0.2	460	152	231	6.5	3.2	0.9	Male	NLP
46	9.4	5.2	268	21	63	6.4	2.8	0.8	Male	LP
54	0.8	0.2	181	35	20	5.5	2.7	0.96	Male	LP
46	0.6	0.2	115	14	11	6.9	3.4	0.9	Male	LP
40	0.9	0.3	293	232	245	6.8	3.1	0.8	Female	LP
58	0.9	0.2	1100	25	36	7.1	3.5	0.9	Male	LP
45	20.2	11.7	188	47	32	5.4	2.3	0.7	Male	LP
50	0.7	0.2	188	12	14	7	3.4	0.9	Male	LP
58	1	0.5	158	37	43	7.2	3.6	1	Male	LP
41	0.9	0.2	201	31	24	7.6	3.8	1	Female	NLP
65	0.8	0.2	162	30	90	3.8	1.4	0.5	Male	LP
33	1.2	0.3	498	28	25	7	3	0.7	Male	LP
40	3.6	1.8	285	50	60	7	2.9	0.7	Male	LP
37	1.8	0.8	145	62	58	5.7	2.9	1	Male	LP
44	1.9	0.6	298	378	602	6.6	3.3	1	Female	LP
30	0.8	0.2	158	25	22	7.9	4.5	1.3	Female	NLP
23	2.3	0.8	509	28	44	6.9	2.9	0.7	Female	NLP
36	2.8	1.5	305	28	76	5.9	2.5	0.7	Male	LP
30	0.7	0.2	262	15	18	9.6	4.7	1.2	Male	LP
42	2.3	1.1	292	29	39	4.1	1.8	0.7	Female	LP
18	0.8	0.2	282	72	140	5.5	2.5	0.8	Male	LP
32	15	8.2	289	58	80	5.3	2.2	0.7	Male	LP
28	0.6	0.1	177	36	29	6.9	4.1	1.4	Male	NLP
48	5.8	2.5	802	133	88	6	2.8	0.8	Male	LP

50	1.7	0.6	430	28	32	6.8	3.5	1	Female	LP
40	14.5	6.4	358	50	75	5.7	2.1	0.5	Male	LP
70	1.4	0.6	146	12	24	6.2	3.8	1.58	Male	NLP
55	0.9	0.2	190	25	28	5.9	2.7	0.8	Male	LP
49	1.1	0.5	159	30	31	7	4.3	1.5	Male	LP
30	1.6	0.4	332	84	139	5.6	2.7	0.9	Male	LP
45	0.8	0.2	165	22	18	8.2	4.1	1	Female	LP
60	5.2	2.4	168	126	202	6.8	2.9	0.7	Male	LP
18	0.8	0.2	228	55	54	6.9	4	1.3	Male	LP
50	0.7	0.1	192	20	41	7.3	3.3	0.8	Female	LP
58	0.8	0.2	180	32	25	8.2	4.4	1.1	Male	NLP
17	0.5	0.1	206	28	21	7.1	4.5	1.7	Female	NLP
70	0.6	0.1	862	76	180	6.3	2.7	0.75	Male	LP
57	1	0.3	187	19	23	5.2	2.9	1.2	Male	NLP
26	1.3	0.4	173	38	62	8	4	1	Male	LP
51	0.8	0.2	367	42	18	5.2	2	0.6	Male	LP
48	0.8	0.2	142	26	25	6	2.6	0.7	Female	LP
54	0.8	0.2	218	20	19	6.3	2.5	0.6	Male	LP
64	0.8	0.2	178	17	18	6.3	3.1	0.9	Female	LP
45	0.7	0.2	164	21	53	4.5	1.4	0.45	Female	NLP
46	0.6	0.2	290	26	21	6	3	1	Male	LP
45	2.5	1.2	163	28	22	7.6	4	1.1	Male	LP
31	0.6	0.1	175	48	34	6	3.7	1.6	Male	LP
38	0.9	0.3	310	15	25	5.5	2.7	1	Male	LP
18	0.8	0.2	199	34	31	6.5	3.5	1.16	Female	NLP

60	8.6	4	298	412	850	7.4	3	0.6	Male	LP
60	0.9	0.3	168	16	24	6.7	3	0.8	Male	LP
50	1.7	0.8	331	36	53	7.3	3.4	0.9	Male	LP
55	75	3.6	332	40	66	6.2	2.5	0.6	Male	LP
42	0.9	0.2	165	26	29	8.5	4.4	1	Female	NLP
13	0.7	0.1	182	24	19	8.9	4.9	1.2	Female	LP
56	1	0.3	195	22	28	5.8	2.6	0.8	Male	NLP
35	1.8	0.6	275	48	178	6.5	3.2	0.9	Male	NLP
38	0.8	0.2	208	25	50	7.1	3.7	1	Male	LP
40	0.6	0.1	171	20	17	5.4	2.5	0.8	Male	LP
60	22.8	12.6	962	53	41	6.9	3.3	0.9	Male	LP
72	0.7	0.1	196	20	35	5.8	2	0.5	Male	LP
43	0.8	0.2	192	29	20	6	2.9	0.9	Male	NLP
42	0.7	0.2	197	64	33	5.8	2.4	0.7	Male	NLP
35	1.6	0.7	157	15	44	5.2	2.5	0.9	Male	LP
30	0.7	0.2	63	31	27	5.8	3.4	1.4	Female	LP
20	16.7	8.4	200	91	101	6.9	3.5	1.02	Female	LP
62	1.8	0.9	224	69	155	8.6	4	0.8	Male	LP
49	0.6	0.1	185	17	26	6.6	2.9	0.7	Female	NLP
50	0.9	0.3	194	190	73	7.5	3.9	1	Male	LP
52	0.6	0.1	194	10	12	6.9	3.3	0.9	Female	NLP
60	5.8	2.7	204	220	400	7	3	0.7	Male	LP
35	0.9	0.3	158	20	16	8	4	1	Female	LP
34	5.9	2.5	290	45	233	5.6	2.7	0.9	Male	LP
51	2.9	1.3	482	22	34	7	2.4	0.5	Male	LP

50	7.3	3.7	92	44	236	6.8	1.6	0.3	Male	LP
48	2.4	1.1	554	141	73	7.5	3.6	0.9	Male	LP
66	0.7	0.2	239	27	26	6.3	3.7	1.4	Male	LP
21	0.6	0.1	186	25	22	6.8	3.4	1	Female	LP
19	0.7	0.2	186	166	397	5.5	3	1.2	Female	LP
18	0.7	0.1	312	308	405	6.9	3.7	1.1	Male	LP
38	3.1	1.6	253	80	406	6.8	3.9	1.3	Male	LP
66	1	0.3	190	30	54	5.3	2.1	0.6	Male	LP
36	0.9	0.1	486	25	34	5.9	2.8	0.9	Male	NLP
17	0.9	0.3	202	22	19	7.4	4.1	1.2	Male	NLP
46	18.4	8.5	450	119	230	7.5	3.3	0.7	Male	LP

Appendix G

Kyphosis (Training and Test)

Training

Age	Number	Start	Kyphosis
36	4	13	absent
42	7	6	present
120	2	13	absent
26	7	13	absent
157	3	13	present
178	4	15	absent
11	3	15	absent
206	4	10	absent
87	4	16	absent
127	4	12	absent
158	5	14	absent
15	5	16	absent
18	4	11	absent
159	4	13	absent
195	2	17	absent
17	4	10	absent
118	4	16	absent
118	3	16	absent
81	4	1	absent
114	7	8	present
130	4	1	present

102	3	13	absent
51	7	9	absent
120	5	8	present
2	3	13	absent
72	5	15	absent
140	4	15	absent
2	2	17	absent
139	10	6	present
9	2	17	absent
68	5	10	absent
177	2	14	absent
121	3	3	present
131	5	13	absent
136	4	15	absent
139	3	10	present
97	3	16	absent
61	4	1	absent
143	9	3	absent
35	3	13	absent
73	5	1	present
91	5	12	present
20	6	9	absent
52	5	6	present
1	3	9	absent
93	3	16	absent

140	5	11	absent
112	3	16	absent
130	5	13	absent
125	2	11	absent
31	3	16	absent
151	2	16	absent
4	3	16	absent
100	3	14	absent

Test

8	3	6	absent
9	5	13	absent
15	7	2	present
131	2	3	absent
96	3	12	present
105	6	5	present
22	2	16	absent
27	4	9	absent
80	5	16	absent
175	5	13	absent
78	6	15	absent
1	3	16	absent
168	3	18	absent
1	4	12	absent
18	5	2	absent
148	3	16	absent

82	5	14	present
59	6	12	present
113	2	16	absent
37	3	16	absent
61	2	17	absent
1	2	16	absent
1	4	15	absent
2	5	1	absent
128	4	5	present
158	3	14	absent
71	3	5	absent