Data mining techniques and breast cancer prediction : A case study of Libya.

ABDULL, Mohamed A. Salem.

REFERENCE

Data Mining Techniques and Breast Cancer Prediction: A Case Study of Libya

By

Mohamed A. Salem Abdull

A thesis submitted in partial fulfilment of the requirement of

Sheffield Hallam University

For the degree of Doctor of philosophy

December 2011

Faculty of Art, computing, Engineering and Sciences

Sheffield Hallam University

DECLARATION

I declare that the work presented in this thesis is my own work, done to the best of my knowledge and effort except as acknowledged in the text, and the work referred to in this thesis has not been submitted, either in whole or in part, for another degree or qualification.

ACKNOWLEDGEMENT

I would like to express many thanks to my supervisors, Dr Patrick Ezepue and Dr. Kassim Mwitondi, for their effort, valuable comments, guidance and assistance during my research; it would have been next to impossible to write this thesis without their help. Their enthusiastic supervision, patience and invaluable technical suggestions throughout this research helped me defeat many crisis situations to finish this dissertation and helped my personal development as a researcher. They both gave me the advantage of their expertise and generosity.

I am also grateful to them for evaluating my work, commenting on my views and helping me understand and develop my ideas. My deepest appreciation goes to Dr Frances Slack, Head of Postgraduate Research Dr Kathy Doherty, and Director of the Cultural, Communication and Computing Research Institute Professor Simeon Yates, for giving me support and solving my problems. Their infinite support for all my needs has been so valuable to me. It is a pleasure to thank all the administrative staff, especially Mrs Tracey Smith, and my colleagues in the Faculty of Arts, Computing Engineering and Sciences (ACES) at Sheffield Hallam University, particularly those in the Culture, Communication and Computing Research Institute (C3Ri). Also, I would like to express my gratitude and appreciation to the Libyan Ministry of Higher Education for the scholarship award and support throughout the time of my studies.

DEDICATION

I owe my deepest gratitude to my parents Abdulsalam and Fatima who gave me their moral support, encouragement and understanding. I would especially like to thank my wife Huwaida as I would never have been able to do this without her. She has given unconditional love and provided me with power, dreams, courage and the determination to finish this thesis. Also, I thank all my family, Fatmha, Abdullsalam, Kawtar, Almotamed, and all my brothers and sisters for giving me their inspiration, patience and tolerance throughout my preoccupation with this work. At the same time, I would like to thank all my friends who have given me their sincere help, suggestions and guidance. I am heartily thankful to my best friend Mahmuod Atayep, who always found time to listen to my problems during my research. He gave me the advice and encouragement to finish this research and is a magnificent support. Also, another special thanks to my best friend Khalid Alzubi from Jordan, who is always available to talk to me and cheer me up when I need him. He gave the motivation to carry on doing my best during this research as he gave the best moral support that I can never forget. Last but not least, thanks are to God for giving me the strength and answering my prayers to finish this research.

ABSTRACT

Different forms of cancer have been widely studied and documented in various studies across the world. However, there have not been many similar studies in the developing countries - particularly those on the African continent (Parkin, *et al.,* 2005). This thesis seeks to uncover the geo-demographic occurrence patterns of the disease by applying three Data mining Techniques, namely Logistic Regression (LR), Neural Networks (NNs) and Decision Trees (DTs), to learn the underlying rules in the overall behaviour of breast cancer. The data, 3,057 observations on 29 variables obtained from four cancer treatment centres in Libya (2004-2008), were interrogated using multiple K-folds cross validation. The predictive strategy yielded a list of breast cancer predictor factors ordered according to their importance in predicting the disease. Comparison between our results and those obtainable from conventional LR, NN and DT models shows that our strategy out-performs the conventional variable selection. It is expected that the findings from this thesis will provide an input into comparative geo-ethnic studies of cancer and provide informed intervention guidelines in the prevention and cure of the disease, not only in Libya but also in other parts of the world.

PUBLICATION

The following paper was produced to publish the concept and result of the work undertaken during the course of this PhD study: Salem, M.A, Mwitondi, K.S (2010); Predicting breast cancer using combined K-fold cross validated decision tree models; The International CODATA Conference 24-27 October 2010 Cape Town, South Africa.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

LT	Logistic Regression
NN	Neural Networks
DT	Decision Trees
TB	Tuberculosis
HIV	Human immunodeficiency virus
AIDS	Acquired immune deficiency syndrome
WHO	World Health Organization
CI	Confidence Interval
OR	Odds Ratio
SES	Socio-Economic Status
RRs	Rate Ratios
X2	Chi-Square
SIR	Standardized Incidence
SMR	Mortality Ratios
BMI	Body Mass Index
HR	Hazard Ratios
OC	Oral Contraceptive
DNA	Deoxyribonucleic Acid
LDA	Linear Discriminant Analysis
PM	Probability of Misclassification
CM	Cost of Misclassification
WHR	waist/hip Ratio
ANNs	Artificial Neural Networks
CART	Classification and Regression Trees
CPs	Complexity parameters

Chapter 1: Introduction

1.1 Introduction

Different forms of cancer have been widely studied and documented in various studies across the world. Ames *et al* (1995), Little (2001) and Jerez-Aragones *et al* (2003) investigated the patterns of breast cancer spread in the United States on the basis of sampled demographic data combined with breast cancer mortality rates among women and just like in some other studies they uncovered geographical, age, gender and racial patterns. However, not as many studies have been carried out in the developing countries. Parkin *et al* (2005) reported that documented cancer statistics are particularly low on the African continent. Hence there is a pressing need for a thorough study of cancer predisposing factors across the continent. Fighting the spread of life-threatening diseases and providing cost effective remedies is a global priority.

This thesis focuses on the prevalence of breast cancer on the African continent and seeks to investigate the geographical patterns of the disease, assess and evaluate current predictive models used in predicting its occurrence and develop and/or enhance breast cancer predicting models. It specifically focuses on understanding the geo-demographic patterns of occurrence and development of the disease as a basis for informed intervention in its prevention and cure. Hence, its main goal is to gain a deeper understanding of the factors causing cancer, occurrence and development of the disease as basis for intervention for prevention and cure. Using domain-partitioning techniques - logistic regression (LT), decision trees (DT) and neural networks (NN), the thesis seeks to attain efficiency in modelling the disease conditions and assessment of the performance of the models on the available data on breast cancer.

1.2 Motivation of research

The study is motivated by the current global situation of breast cancer and the methodologies used in collecting, analysing and sharing data and knowledge relating to the disease. More specifically, it is motivated by the scope, geographical and ethnic aspects of the disease and how much is known about it. Kerr *et al* (2007) report that the disease worldwide is responsible for more than 7 million deaths yearly; more than malaria, TB and HIV/AIDS combined. In the developing world, the number of new cancer cases will increase significantly over the next ten years. Also, Kerr et al state that African countries will have over a million new cancer cases a year and they are the least able of all developing countries to cope, having fewer cancer care services. By 2020

there would be 15 million new cases of cancer every year, 70% of which will be in developing countries, where governments are least prepared to face cancer and where survival rates are often less than half those of more developed countries. (Gonzaga, 2010)

Worldwide, more than one million new cases of female breast cancer are diagnosed each year. It is the most commonly occurring neoplasm in women, accounting for over one-fifth of the estimated annual 4.7 million cancer diagnoses in females, and the second most common tumour, after lung cancer, in both sexes. It is also the most common female cancer in both developing and developed countries, with most (55%) occurring in the latter regions, where age-standardised rates are three times higher than in developing areas (Bray *et al* 2004).

According to Kruger and Appelstaedt (2007), breast cancer is rising within the lower socio-economic groups in Africa and may in the medium term become a problem for the African population. Although treatment is often considered to be connected to primary prevention, it has been estimated that between 2000 and 2020 approximately 10 million patients will die of cancer in Africa. Mortality rates are higher in Africa than in richer world regions and improved access to known effective therapy, efficiently delivered, would, therefore, save lives. They also reports that breast cancer also occurs in younger African women more than in other parts of the world.

Various methods have been used in analysing breast cancer data. For instance, Gilliland *et al* (2001) use logistic regression to investigate breast cancer risk in Hispanic and non-Hispanic white women in the United States, and they found that the effect of physical activity were larger among premenopausal Hispanic than non-Hispanic white women. Also, they found that the overall protective effects of physical activity were larger in Hispanic than non-Hispanic white women.

Expectations are that findings from this study will help decision and policy makers across the African continent and beyond. The review of the modelling techniques is expected to add to a portfolio of tools and techniques used in modelling the disease worldwide. It follows therefore that further research based on accurate data collection and analysis is still required.

Our study has some methodological limitations. The quantitative approach uses a questionnaire; it provides a wide scope for investigation, but perhaps less so for detailed

explanation, whereas a qualitative focus would be narrower but more exhaustive. Also, Libya is a country huge in area, and centres for treatment are situated in the north; patients only come to centres for treatment and do not stay there. Thus it is very difficult for the researcher to interview them. Information was taken from their folders, and if any data were missing, we posted a questionnaire or called them. Every effort has been made to ensure the inclusion of all relevant information regarding the patients' cases and control in this research.

1.3 Research questions and study objectives

This study sets off from the current state of knowledge and practice in identifying factors causing breast cancer seeks to answer the following research questions:

7. ***Are breast cancer predictor factors the same across the world?*** To answer this question we look at the geo-demographics of breast cancer predictor factors with particular attention to the African continent (using Libya as a case study).
2. ***How can we enhance breast cancer predicting models?*** To answer *this question we explore the performance of three common models: LR, DT and NN.*

The study's main objectives of can be summarised as follows:

1. ***To explore the theoretical and practical aspects of the literature based on previous cancer studies in order to understand the range of models used by previous breast cancer researchers and the factors associated with the disease.***
2. ***To use insight gained from objective 1 to develop predictive models appropriate for estimating risks of developing breast cancer by women with given characteristics, determining risk of re-occurrence of the disease in a patient who has recovered with treatment, and other aspects of the disease.***
3. ***To provide breast cancer findings from the same cultural background, Libya, as a comparative input into the global geographical distribution of breast cancer.***

1.4 Contribution to knowledge and benefits to society

The research shows how the models are applied to breast cancer data from Africa. Since there is no evidence in the literature of these models being used in Africa, the modelling results from this research will be a significant contribution to the statistical analysis of cancer data in Africa. Work on Objectives will contribute useful new knowledge on the structure of breast cancer datasets from different regions of Libya, which will support further research and modelling of the disease throughout Africa and in other developing

countries. The geographical comparisons of data/results from the different regions are also further contributions to knowledge.

In addition, our research would be useful to professionals and researchers in the field of breast cancer research worldwide, especially in non-developed countries.

It is expected that the results will help determine appropriate approaches for preventing and treating the disease in different regions of Africa for different groups of patients. More specifically, it will facilitate the selection of appropriate models and techniques for analysing the data on breast cancer to be used in future studies. The main government which would benefit from our research is Libya, since our study would provide the policy makers with important information to enable them to design better strategies to fight the disease. On the other hand, The World Health Organization (WHO) would definitely use the results of our research and add it to the database of breast cancer research in the African continent. Moreover, our research will be a reference to those researchers who are interested in investigating the effect of geographical, environmental, and ethnic risk factors on developing breast cancer disease. Finally, the research results could be presented in many international conferences, or as published papers in professional Journals.

1.5 Structure of thesis

The thesis is divided into seven chapters. The first chapter presents an introduction to our study and provides the motivation for carrying out the study on breast cancer. The chapter also clarifies the main objectives of the study and the research questions. Chapter two explores the literature and reviews previous work done by other researchers in modelling breast cancer. Chapter three describes the research methodology adopted in the study, the modelling approaches and techniques used - namely, the Bayesian, cross-validation and data mining techniques such as Logistic Regression, Neural Network and Decision Trees - and outlines our main strategy for variable selection. Chapter four presents data analysis - exploratory and strategic, Chapter five devoted to further Analysis of data: 2-stage modelling and finally, chapter seven provides summarises and evaluates the work; outlines its contribution and limitation and makes recommendations for further work.

Chapter 2 : Literature Review

2.1 Introduction

The main purpose of the current chapter is to explore the background of factors predicting breast cancer, which in turn will provide insights into the prevention and cure of the disease. Consequently, this demonstrates a plan to avoid these factors or at least to minimize them as much as possible. This explains the importance of exploring the factors causing breast cancer. If we can precisely detect these factors, then we will be able to fight the disease effectively. There is another reason behind exploring risk factors causing breast cancer, which is a life threatening disease. Breast cancer, as the highest occurring cancer among women, is a major health burden worldwide, causing over one million of the estimated 10 million cases diagnosed worldwide each year in both sexes, and the primary cause of cancer deaths among women globally. Due to it, 375,000 deaths occurred in the year 2000 (Bray *et al,* 2004). Numerous breast cancer risk factors have been widely studied in different combinations, as summarised below.

2.2 Demographical risk factors

Of all the risk factors, gender has emerged as the most significant predictor of breast cancer; although men can get breast cancer, women account for more than 99% of all breast cancer cases.1 Amir *et al* (1996) reported that In an African population, the occurrence of this cancer is high. The male/female ratio in Tanzania is 1:14 (0.071). This narrow ratio does not differ significantly in the majority of sub-Saharan African countries, the overall ratio being 0.0143 (Cl = 0.0317-0.877).2 Another important predictor of breast cancer is age. According to Katapodi *et al.* (2005), the lifetime risk of a woman getting breast cancer is usually estimated at one in eight, but this probability does increase with age. Further, the difference in probability levels over the lifespan range is from one in 233 between the ages of 30 and 39 to one in 27 between the ages of 60 and 69 (Radice *et al,* 2003).

1 http://www.ucsfhealth.org/adult/medical_services/cancer/breast/risks.html

2 http://www.ncbi.nlm.nih.gov/pubmed/8698019

In the Nightingale Centre, University Hospital of South Manchester, between August 1993 and October 1994, a study by Warwick *et al.* (2002) included the age factor in all multivariate stepwise ordered and unordered logistic models to identify the risk factors which remained significant after controlling for other factors; it was found that increase in age by a year results in a 7% reduction in breast cancer risk. In another study, Parkin *et al* (2005) reported that an average European woman aged about 25 years had a 1 in 15000 chance of developing breast cancer while those 40 years old had a risk of 1 in 200. For women aged about 50 years the chance was 1 in 50, and for those about 80 years it was 1 in 11.

Various researches suggest that physical activities or labour can change the menstrual cycle which results in a change in hormonal level (Stemfeld *et al,* 2002). The risk of breast cancer may be reduced due to the physical activities performed by women altering their menstrual cycle patterns and thus altering the production of ovarian hormones. One research carried out by Bernstein *et al* (1994) used univariate and multivariate conditional logistic regression methods to analyse the data from case-control studies for all white female residents of Los Angeles County who were bom in the United States, Canada, or Europe. The results showed that the average age of case patients at diagnosis was 36 years and the average age of control subjects was 36 years old. Also, cumulative exposure to ovarian hormones is a determinant of breast cancer risk.

Abdalkader *et al* (2008) used Tow way ANOVA to identify the effects of night shift working on Jordanian nurses in critical care units. They found that irregular working hours, especially during the night-time, and long night shifts result in changes in psychological and physiological factors and increase the chances of breast cancer in women: the same research does not apply to men. They especially suggested that women with irregular working hours are at increased risk for breast cancer since work that requires the use of artificial light (in the evening, night or early in the morning) leads to the suppression of pineal secretion of melatonin, which may induce continuous production of estrogens involved in breast carcinogenesis.

Epidemiologic studies show that occurrence of breast cancer is higher in poor, low or no education, and low socioeconomic factors. The origin of these differences is still unknown. However, survival of breast cancer patients depends on education level and socioeconomic factors. Improvement in lifestyle, treatment and knowledge about breast

cancer, and issues related to it can greatly change the statistics in a short period of time. Webster *et al.* (2008) in a comparison study for cases and control patients diagnosed between 1987 and 1993 on Cape Cod, Massachusetts (USA) using LR to calculate the odds ratio (OR) found that women with the highest education were at greater risk of developing breast cancer in both 1980 and 1990 [(OR) = 1.17 and 1.19, respectively]. Similarly, women living in the highest SES communities in 1990 had a greater risk (OR = 1.30).

Eaker *et al.* (2009) conducted a study to assess the presence of social differences in breast cancer survival among patients managed within a national health care system, and whether any such gradients could be explained by disparities in tumour characteristics and management. The research was carried out in Sweden; individual data from several different population-based registers were collected and examined by using both Wald test and likelihood ratio tests to assess the significance of the variables of interest in women with different educational backgrounds. It was concluded that the risk of dying of breast cancer was 35% lower among women with high compared to low education. Compared to women with high education, a lower percentage of women with low education had been investigated. These results suggest that breast cancer risks and its treatment can produce better results in an educated than an uneducated population. Awareness and precaution can greatly reduce the incidence of breast cancer cases.

Research associated with breast cancer risk factors shows that there may be a link between breast cancer and marital status. That encouraged Ebrahimi *et al.* (2002) to conduct a controlled research in Iranian women. During their study demographic data and risk factor related information were collected using a short structured questionnaire. Univariate (LR) analysis was performed to calculate the ORs and to examine the predictive effect of each factor on risk for breast cancer. The significant factors were carried forward and were entered into multivariate LR analysis, which showed that unmarried women are at higher risk than married (OR 2.87, 95% Cl 1.13-7.30). Nulliparous married women were found to have a similar increased risk of breast cancer as compared to parous women of the same age.

Breast cancer is one of the few cancers to have a higher incidence among the more affluent social classes. To evaluate the relationship between socio-economic status (SES) and breast cancer incidence in California for four race/ethnic groups, Yost *et al* (2001) used principal component analysis to create an SES index using 1990 census

data. Untreated cases were randomly allocated to census block groups within their county of residence. A total of 97,227 female breast cancer cases diagnosed in California between 1988 and 1992 were evaluated. Incidence rates and rate ratios (RRs) were estimated and a Chi-square (x2) test for trend across SES levels was performed. The results show that SES was positively related to breast cancer incidence and this effect was stronger for Hispanics and Asian/others than for whites and blacks. The authors concluded that their results are consistent with similar findings for the Los Angeles area, but differ from previous results for the San Francisco Bay area.

A considerable socioeconomic difference prevailed in the burden of breast cancer among Danish women between 1970-1995 Dano *et al* (2003) used a Poisson distribution and also calculated the standardized incidence (SIR) and standardized mortality ratios (SMR) values for married, economically active women on their own. The results show that academics had the highest risk and women working in agriculture had the lowest risk.

Also examining whether women living in such communities remained at greater risk of breast cancer after controlling for individual education and other known individual-level risk factors, Robert *et al* (2004) collected data from a population-based, breast cancer case-control study conducted in Wisconsin United States from 1988 to 1995.
The authors, after using multilevel logistic regression models, concluded that women living in the highest SES communities had greater odds of having breast cancer than women living in the lowest SES communities (1.20; 95% confidence interval = 1.05-1.37). Similarly, the odds were greater for women in urban versus rural communities (1.17; 1.06-1.28).

It is logical to have a very strong relation between socio-economic status and breast cancer, since high socio-economic status is related to a modem life style and having good welfare standards. The lower socio-economic status of most African countries compared to developed countries leads to a lack of screening, medication, specialists, and lack of adequate databases and registers to update the records of cancer patients.

2.3 Controlled factors

The relation between body size and breast cancer risk has been the subject of numerous investigations. In a case-control study among Asian-American women living in the western United States, Van den Brandit *et al* (2000) used conditional logistic regression

with the use of SAS software. They found that the association of weight with breast cancer depended on menopausal status. There was an inverse association in premenopausal women, concentrated in the top weight categories. For postmenopausal women, a positive association was found. Body mass index (BMI, defined as weight (kg)/height^{9} (m^{9})) showed a significant inverse association with breast cancer risk among premenopausal women. On the other hand, the association with BMI among postmenopausal women was significantly positive.

In a population-based case-control study of 479 women with incident primary breast cancer and 435 controls from western Washington State, unconditional logistic regression analysis was performed using SPSS Software; odds ratios (OR) and 95% Cl were calculated to estimate the relative risk of breast cancer for the various factors. Li *et al.* (2000) found that when BMI was broken into quartiles, women in the highest BMI quartile had an increased risk of breast cancer when compared to women in the lowest quartile (OR = 1.5, 95% Cl, 1.1-2.3).

Adebamowo *et al.* (2003) used LR to examine data concerning the relationship between waist-hip ratio and the risk of breast cancer in an urban Nigerian population, demonstrating - like similar studies - a positive association between obesity and the risk of breast cancer among postmenopausal African women. Being overweight is associated with a doubling of the risk of breast cancer in postmenopausal women whereas amongst premenopausal women obesity is associated with reduced breast cancer incidence.3

Obesity increases the risk of postmenopausal breast cancer by up to 30%. A case - control study of Dutch women found a positive relati onship only in postmenopausal women (Stephenson and Rose 2003). Also, population -based cohort studies on obesity among older postmenopausal women noted an increased risk of breast cancer with increasing BMI (Sweeney *et al.,* 2004; Krebs *et al.,* 2006).

We can see that modem lifestyle badly affects bodyweight, due to the consumption of fast food and many other related factors, such as lack of manual work and the near-entire use of mechanised products that reduce physical exertion. On the other hand, African women do manual and physical jobs which reduce their possibility of being overweight.

3http://www.tiscali.co.uk/lifestyle/healthfitness/health_advice/netdoctor/archive/000092.html

The relationship between diet and breast cancer risk is often divided into specific items of consumption (Zografos *et a l,* 2004). Hanf and Gonder (2005) aptly noted that "there are as many publications on cancer and its relation to nutrition as there are different foods to savour". A strong positive relationship has been observed in red meat consumption. Breast cancer risk was especially obvious among premenopausal women who were estrogen and progesterone receptor positive and who consumed more than 1.5 servings per day of red meat (Cho *et al,* 2006). Meanwhile Missmer *et al.* (2002) estimated rate ratios (RR) by conditional logistic regression models using SAS Software combining the primary data from eight prospective cohort studies from North America and Western Europe. They found no statistically significant association between red meat consumption and breast cancer risk.

Engeset *et al.* (2006) collected data from 23 centres in ten European countries (Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden and the United Kingdom). Hazard ratios (HR) and their 95% Cl were estimated in their study using the Cox proportional regression model. They found no relationship between fish consumption and breast cancer risk among over 300,000 women. Similar results were found in a case-control study of over 2,000 Swedish breast cancer cases in that type of fish and amount consumed per week were not shown to have a statistically significant relationship with breast cancer risk (Terry *et al,* 2002). The majority of studies exploring the relationship between fruit and vegetable consumption indicate a protective effect against breast cancer (Moorman and Terry 2004; Zografos *et al.,* 2004).

With respect to drinking coffee and tea, Baker *et al.* (2006) found an increase in lobular breast cancer among premenopausal women who drank a cup of coffee or less a day, and a decrease in lobular cancer among black tea drinkers. A meta-analysis of three Japanese cohort studies and one case-control study in California found a negative association between high levels of green tea consumption (at least five cups a day) and breast cancer risk (Sun *et al.,* 2006).In the literature many researchers have investigated the relationship between physical activity and breast cancer risk. Gilliland *et al* (2001) undertook a controlled population-based case study among Hispanic and non-Hispanic women to estimate the relative risk of breast cancer for levels of physical activity.

They used Conditional LR which, conditioned on the frequency-matched variables (three age groups, geographical district and ethnicity), was used to compute odds ratios

and 95% CIs. Gilliland *et al* found that the effects of physical activity in premenopausal white women are greater for Hispanics than for non-Hispanics, while only their non-Hispanic postmenopausal counterparts were thus protected.

In another case-control study of three racial/ethnic groups in the San Francisco area, John *et al.* (2003) used an unconditional logistic regression model to calculate ORs and 95% CIs as an estimate of the relative risk associated with various physical activity measures. They declared that an increase in lifetime physical activity at both low and high levels was associated with a lower risk of breast cancer in both pre- and post-menopausal women.

On the other hand, in three large cohort studies Colditz *et al* (2003), Margolis *et al.* (2005) and Mertens *et al* (2006) found no relationship between physical activity and breast cancer in pre- or post-menopausal women. When compared to the studies where an inverse relationship between physical activity and breast cancer risk was observed, these three studies all relied on self-reported data of physical activity at specific points in time as opposed to an assessment of lifetime physical activity. Zografos *et al* (2004) and Mertens *et al* (2006) have shown an inverse relationship between physical activity and breast cancer risk.

In another case-control study of 4,538 breast cancer cases of Black and White women in the United States, Bernstein *et al* (2005) used unconditional logistic regression; they noted that an increase in lifetime physical activity was associated with a decreased risk of breast cancer.

The relationship between breastfeeding and breast cancer risk is most often noted in the inverse, such that the longer the duration, the lower the risk. In a reanalysis of data from 47 epidemiological studies in 30 countries of almost 150,000 women, the relative risk is reduced by 4.3% for each year breastfeeding (Collaborative Group on Hormonal Factors and Breast Cancer, 2002).

A case-control study of 404 breast cancer cases and an equal number of controls in Shandong Province, China, (Zheng *et al.,* 2000), using both conditional and unconditional logistic regression, found that women who breastfed for more than 24 months per child had a lower odds ratio for breast cancer than those women who only breastfed for 1 to 6 months per child (OR = .46, 95% Cl, 0.27 to 0.78). Similar results were found by using multivariate logistic regression in a case-control study of 349

Mexican women, but breastfeeding the firstborn baby was found to have an even greater protective effective against breast cancer (Romieu *et al.,* 1996).

The lack of long-term breastfeeding practices in developed countries may account for the higher incidence rates of breast cancer (Collaborative Group on Hormonal Factors and Breast Cancer, 2002). African women in general encourage breastfeeding more than others, due to religious and cultural beliefs. This implies a reduced risk of getting breast cancer in African women. On the other hand, the majority of women in developed countries have fewer children and do not encourage breastfeeding due to a fast and busy lifestyle that encourages them to use alternative artificial milk products.

A controlled case study of Thai women was carried out by Susan Jordan *et al.* (2009). Odds ratios and 95% confidence intervals were calculated using conditional logistic regression. They found no significant difference in the occurrence of breast cancer between women who had children and those who had none (OR=0.9, 95% Cl 0.5-1.7 for ever versus never having children) nor was there a significant relation with number of births or the age of a woman when her first child was bom (OR = 1.0, 95% Cl 0.3-3.1 for those aged 30 or more at first birth versus those aged 25 or less).

Breast cancer is another controversial issue associated with OC4 use. Epidemiologic studies have generally not shown any relationship between oral contraception and occurrence of breast cancer, whereas other studies have suggested that there may be an increased risk of developing breast cancer in women who use oral contraceptive pills when they are less than 35 years of age. In a case-control study of women aged 20 to 44 years, 1648 cases of breast cancer and 1505 control subjects were identified. Oral contraceptive pill use for 6 months to 5 years among women less than 45 years of age was associated with a 1.3 relative risk for breast cancer development (95% Cl, 1.1 to 1.5). This risk was increased to 1.7 (95% Cl, 1.2 to 2.6) in oral contraceptive users less than 35 years of age, with the risk increasing to 2.2 (95% Cl, 1.2 to 4.1) in women using the pill for more than 10 years and 3.1 (95% Cl, 1.4 to 6.7) in women who also began using the pill before age 18 (Brinton *et al,* 1995).

2.4 Uncontrolled factors

Taller women have an increased risk of breast cancer. Ziegler *et al.* (1996) conducted a

case-control study of breast cancer among women of Chinese, Japanese and Filipino ethnicities, aged 20-55 years, living in San Francisco-Oakland, Los Angeles, and Oahu during the period from 1983 to 1987. Logistic regression was performed to obtain maximum likelihood estimates of the odds ratios and 95% confidence intervals (CIs).The results showed that median height of the control of Asian-American women was 62 inches (1.57 m); the median heights for the Chinese, Japanese, and Filipino control were 62,61 (1.55 m), and 61 inches, respectively. Also they reported that relative risk (RR) of breast cancer increased steadily with height; women with a height of 66 inches or more (>1.66 m) were at twice the risk (RR = 2.01; 95% Cl = 1.16-3.49) of women height of 59 inches or less (<1.51 m). The influence of height on breast cancer risk was similar in premenopausal and postmenopausal women.

LR models were used to calculate the ORs by Furberg *et al* (2002), who examined the potential etiologic heterogeneity of breast cancer by seeking to determine the existence of any association between cigarette smoking and exposure to low dose ionizing radiation and the disease by conducting a case-control study of 861 African-American and white women aged between 20 and 74. The authors found that the relationship between radiation dose and breast cancer risk can be described by a straight line, which implies that no matter how low the dose, there is some small risk associated with exposure.

Meanwhile, using LR, Gilliland *et al.* (2001) reveal no clear links between exposure to low dose ionizing radiation on the one hand and breast cancer on the other. Laboratory studies have, however, shown that ionized radiation causes damage to DNA, which can potentially increase the chances of breast cancer. The evidence for this comes from many different sources, including studies of atomic bomb survivors in Japan, people exposed during the Chernobyl nuclear accident, people treated with high doses of radiation for cancer and other conditions, and people exposed to high levels of radiation at work, such as uranium miners.5 The chance that women would be exposed to radiation and pollution in developed countries is much higher when compared to such exposure for African women. This can be attributed to the indus trial lifestyle in developed countries with a polluted atmosphere full of carbon dioxide and radiation, compared to the virgin African environment. Bernstein *et al* (1993), Rautalahti *et al*

5http://www.cancer.org/Cancer/CancerCauses/OtherCarcinogens/MedicalTreatments/radiation-exposure-and-cancer

(1993), Nagata *et al* (1995) and Talamini *et al* (1996) state that early age at menarche and late age at menopause have been shown to be risk factors for breast cancer. Titus-Emstoff *et al.* (1998), in a case-control study of 6,888 breast cancer cases and 9529 control women from Wisconsin, Western Massachusetts, Maine, and New Hampshire, using conditional logistic regression models, found that age at menarche of 15 years or older did reduce the risk of breast cancer when compared to a menarche age of 13 years. The protective effect was stronger among premenopausal women, especially premenopausal women who had experienced irregular menstrual cycles up to 5 years after their first menstrual period.

Tavani *et al,* (1999) used logistic regression to investigate the breast cancer risk in women younger than 40 using data from two case-control studies conducted in Italy between 1983 and 1994. Breast cancer was historically confirmed in 579 of the women, while the control numbered 668 women. They found that the risk of breast cancer is inversely related to age at menarche, with a multivariate OR of 0.53, 95% Cl 0.31-0.89 for women reporting menarche at the age of >15 years compared with <12 years. The relationship between age at first birth and breast cancer was first recorded over 30 years ago MacMahon *et al* (1982), as cited in (Leon, 1989). Having no children and being older at the time of the first birth both increase the lifetime incidence of breast cancer. The risk of breast cancer in women who have their first child after the age of 30 is about twice that of women having their first child before the age of 20. The highest risk group are those who have their first child after the age of 35; these women have an even higher risk than women who have no children.6

Wohlfahrt *et al* (2001) in a case-control study of 13,049 breast cancer cases in Danish women bom between 1935 and 1978 using Log-linear Poisson regression models, found that subsequent births at an earlier age also reduce the risk of breast cancer. Since African women in general have more children than other women, and first give birth while they are very young, they are more likely to have a lower incidence of breast cancer than women from, say, developed countries.

An increasing number of women opt to undergo abortion. Has abortion or miscarriage a direct or indirect link with breast cancer? It was always contradicted that spontaneous abortion has any role in developing breast cancer or could act as a risk factor in

6 http://www.netdoctor.co.uk/diseases/facts/breastcancer.htm

developing breast cancer. The scientific evidence does not support the notion that abortion of any kind raises the risk of breast cancer or any other type of cancer." Also, more rigorous recent studies demonstrate no causal relationship between ind uced abortion and a subsequent increase in breast cancer risk (Committee on Gynecologic Practice, 2009).

2.5 Health condition

An increased risk of breast cancer in women with a family history of breast cancer has been demonstrated by many studies using a variety of study designs, for example, Sattin *et al.* (1985) and Pharoah *et al.* (1997). To investigate the association between family history and breast cancer in Mexican women, Calderon *et al* (2000) used the data obtained from a case-control study of 151 breast cancer cases and 235 controls. By using a multiple logistic regressions model to analyse the data, they found a clear association between family and breast cancer.

Based on combined data from 52 epidemiological studies including 58209 women with breast cancer and 101986 controls using conditional logistic regression, the Collaborative Group on Hormonal Factors in Breast Cancer (2001) reported that a woman with one affected first degree relative (mother or sister) had approximately double the risk of breast cancer of a woman with no family history of the disease. If two first-degree relatives developed the disease before the age of 45 years, then a woman's chance of developing breast cancer is four times greater than normal. In the Iranian case study mentioned above, Ebrahimi et al. (2002) declared that breast cancer risk was significantly greater in women with a family history of the disease (OR 2.87, 95% Cl 1.13-7.30).

Mutations of two genes known as BRCA1 and BRCA 2 have long been known to result in higher risks of breast and ovarian cancer in women. Scientists have also recently found that men with certain mutations of these two genes may have an increased risk of early-onset prostate cancer.8 Sasco et *al.* (1993) note a positive correlation exists between male breast cancer and prostate cancer. Hereditary breast cancer accounts for

7http://www.cancer.org/Cancer/BreastCancer/MoreInformation/is-abortion-linked-to-breast-cancer

8 http://prostatecancer.about.eom/od/riskfactors/a/prostatecancerbreastcancerlink.htm

up to 5-10% of all breast carcinomas. BRCA1 and BRCA2 are responsible for about 16% of the familial risk of breast cancer (Tan *et al.,* 2008).

2.6 Geographical Factors

Identifying the factors behind the disease remains of paramount importance. There is evidence that geographical patterns have emerged for particular types of cancer. Gomez-Ruiz *et al* (2004) used Neural Network (NN) to investigate the patterns of breast cancer spread in the United States on the basis of sampled demographic data combined with breast cancer mortality rates among women from all 244 counties in eleven north-eastern states over the period 1988-1992. They concluded that there were higher breast cancer mortality rates in the north-eastern part of the country than in the District of Columbia.

Bray *et al* (2004) reviewed the descriptive epidemiology of the disease, focusing on some of the key elements of the geographical and temporal variations in incidence and mortality in many world regions. They declared that incidence of breast cancer in the USA and Canada is broadly similar to that in European countries; the incidence in New South Wales (representing about one-third of Australian women) increased steadily from the early to mid-1980s, and by 1995 was nearly 50% higher than in 1983. In New Zealand there were steady increases in both Maori and non-Maori incidence rates from 1978-92. In Denmark, both incidence and mortality are declining in young women, and strong cohort effects are observed, with decreasing rates in women bom in successive generations after 1940.

Meanwhile, in Finland, such a reduction in mortality rates has not occurred. Some recent decreases in mortality have been observed in several countries without national screening programmes, although these tend to be confined mainly to younger age groups. Mortality is increasing in several eastern European or former Soviet countries characterised by relatively low rates in the past, such as the Russian Federation, Estonia, Romania and Hungary. Although breast cancer remains relatively rare in Japan, incidence and mortality have been rising quite rapidly; this is consistent with increasing risk in successive generations of women. In the African continent incidence increased in Ibadan, Nigeria and in Kampala, Uganda between the 1960s and the late 1990s.

In Malaysia, where the mortality rate for this disease rose from 0.61 per 100,000 women in 1983 to 1.8 in 1992, Norsa'adah *et al* (2005) conducted a matched case-control study

on data obtained from 147 histologically confirmed breast cancer patients and the same number of non-breast cancer patients with the same spread of age and ethnicity, excluding those with malignant tumours and gynaecological, hormonal or endocrine problems. Simple and multiple conditional LRs were used to analyse the data. Significant risk factors of breast cancer disclosed in their studies include nulliparity, overweight/obesity, family history of breast cancer, and the use of OC. Nulliparity, obesity and family history of breast cancer are well-established risk factors for breast cancer, while the association of OC with breast cancer is still controversial. The study reconfirmed that similar risk factors identified in Western populations were associated with the occurrence of breast cancer in Kelantan Malaysia.

Fregene *et al.* (2005) investigated how breast cancer in Sub-Saharan Africa relates to breast cancer in African-American Women. The results show that Women from sub-Saharan Africa were found to have a low incidence of breast cancer than African-American Women.

Even though data were not available in Northern Africa, El Mistiri *et al.* (2007) presented the first data collected and analyzed in Libya. They collected and analyzed data by the Benghazi Cancer Registry. In 2003, a total of 997 cases of primary cancer were registered among Libyan people. Among females, 26% were breast cancer. The study confirmed that breast cancer incidence is much lower than in western countries.

2.7 Summary

Breast cancer is the uncontrolled growth of cells in the breast and is one of the greatest risks to human health. It mostly affects women, but men also suffer from the disease. Early detection of breast cancer can reduce the rate of mortality. There are different factors which can indicate the occurrence of breast cancer. By close study it has been revealed that early detection is the only way to minimize the effects of breast cancer. The most important factor associated with breast cancer is family history. Other major risk factors which can contribute towards the occurrence of breast cancer are demographics, food, environment, health condition, marital status, breast feeding, menopause, menarche, number of children, and age. The ratio of breast cancer in different regions differs based on different factors. Similarly, mortality rates are different for different regions. In advanced countries the mortality rate is lower than in underdeveloped countries. Thus, the first objective of the study(7o ***explore the theoretical and practical aspects of the literature based on previous cancer studies in***

order to understand the range of models used by previous breast cancer researchers and the factors associated with the disease) is achieved.

From the literature we found many studies have been made across the world to identify the key factors associated with breast cancer, but not many studies have been carried out in Africa (Libya). There is gap in knowledge about the disease in Africa. Our aim is to fill this gap and provide more information about factors related to the disease to help policy makers develop suitable policies to control the disease and to provide relevant information at the right time. This will also be beneficial for the WHO to add to their database relating to developing regions. Also, it was found from the literature search that the three data mining techniques have not previously been used to analyse data from African regions.

In the following chapter we will explore the theory behind the three data mining techniques LR, NN and DT, to identify the most suitable technique.

Chapter 3: Methodology

3.1 Introduction

This chapter focuses on the study's methodology - the collection of tools and techniques used to provide a geographical comparison of breast cancer based on an application of data mining techniques in identifying the factors predictor breast cancer in Libya. The chapter highlights three key aspects: Theoretical overview of predictive model, modelling techniques and the study's strategy.

3.2 Theoretical overview of predictive model

Predictive modelling describes an analytical process used to generate a data-driven model of the future behaviour of a particular phenomenon or its final outcome (Tang, 2009). In most real-life problems predictive models are used to discriminate between groups. A typical example of such models is Fisher's linear discriminant function (Fisher, 1936) also known by such writers as Klecka (1980) as Linear Discriminant Analysis (LDA). The following paragraphs briefly discuss the fundamentals of group discrimination based on LDA.

LDA represents a typical approach to the process of discriminating between groups. The method is used to determine which variables discriminate between two or more naturally occurring groups, such two groups of patients. This is a classical classification problem involving a decision being made as to whether a particular patient belongs to a particular group. The problem is thus reduced to the allocation of each of the patients into one of the two groups. For example, a cancer specialist may record different variables (age, sex, family history, first birth, ethnicity, place of birth and so on) relating to patients' backgrounds in order to learn which variables best predict whether a patient is likely to develop breast cancer (Group 1) or not to develop it (Group 2). The groups' numbers need not be equal.

This procedure can be illustrated by an example involving predicting the occurrence of breast cancer among women with particular characteristics. If it is known a priori that 40 per cent of women with the same characteristics develop the disease, and the remainder do not develop the disease, then it may be assumed that any new dataset with similar characteristics will have priors nx- 0.4 and $tt_2 = 0.6$. If it also known that the data in the two groups are distributed according to the densities [/,(*) and $f_2(x)$], the resulting

prediction will be based on the two priors and the two densities shown in Figure 3-1, giving the probability of membership to each of the groups as:

Allocated to Group 1 if

$$\frac{*1\ M\ x)}{r} > \frac{x\ J\ i\ (*)}{—} — \quad \text{Equation 3-1}$$

Allocated to Group 2 if:

$$\frac{T\ r^{*\ i/i(*)}\ p\ r''}{^{\wedge}\ l/l\ W\ +\ ^{\wedge}\ /i\ W} < \frac{r^{\wedge 2/2\ W}\ r''}{*\ 1/1\ W\ +)r\ 2/!\ W} \quad \text{Equation 3-2}$$

Alternatively, if tt,/, (x) = ?r2f_2(x), allocation is to one of the groups by the same kind of a random rule (Group 1 or Group 2). Such a rule is associated with a total prediction error, £ = + e_2 where corresponds to the incorrect diagnosis of breast cancer in a patient who does not have the disease, and s_2 corresponds to the incorrect clearing of a patient with breast cancer. Clearly, both £± and s_2 are dependent on the priors and the densities - in other words, the probability of observing cases from one of the groups from the viewpoint of another group depends on the error proportions falling under each of the two groups. We can then compute the total probability of misclassification as shown in Equation 3-3

$$PM = \underset{e\ i\ K\ 2\ +\ e\ 2\ Ttl}{£JII2-} + \underset{EiTt2\ +\ E\ 2\ Tti}{-\ Z2U} \quad \text{Equation 3-3}$$

The first additive component of Equation 3-3 gives the probability of observing cases from Group 2 from the standpoint of Group 1, and vice versa for the second component. The parameters and densities used in the formulation of the above rules and probabilities are graphically summarised in Figure 3-1.

Figure 3-1: A graphical illustration of a two-case discrimination rule

The conditional probability of correctly classifying the observation (patients) in Group 1 (breast cancer) is given by:

$$/^{c} \quad /j(x)dx \qquad \text{Equation 3-4}$$

And that of correctly classifying the equivalent observation in Group 2 (no breast cancer) is given by:

$$\boldsymbol{n}_2 = /\mathcal{C}^{\circ} / \mathbf{2}(x)dx \qquad \text{Equation 3-5}$$

As noted above, the overall misclassification error (£) lies in the two tails, and each of the two integrals includes cases from the lower tails of the other integral and excludes cases from its own lower tails. The ultimate goal of this work will be to minimise this error. That is, in an extreme case of success, each of the integrals should exclude zero cases from the lower tails of the other and include all cases from its own lower tails.

There are two main issues of concern here. Firstly, in order to apply the group membership above, the densities and the priors must be known. This is unfortunately often not the case. Secondly, the three equations (Equation 3-1, Equation 3-2 and Equation 3-3) above do not consider the misclassification cost - that is, the repercussions that a person such as a medical doctor may associate with each of the incorrect diagnoses. For instance, a doctor may be more averse to incorrectly making a negative breast cancer diagnosis than vice versa. In this case, the errors above will be weighted by the corresponding proportions as shown in Equation 3-6.where C represents the cost

of predicting "no breast cancer" when the patient actually has the disease, while C_2 is the cost of predicting "breast cancer" when the patient does not have it, as shown in Table 3-1.Consequently, the prediction error in Equation 3-6 is adjusted as the cost of misclassification, as shown.

$$\mathbf{CM} = \frac{\varepsilon_1 \pi_2 C_2}{\varepsilon_1 \pi_2 + \varepsilon_2 \pi_1} + \frac{\varepsilon_2 \pi_1 C_1}{\varepsilon_1 \pi_2 + \varepsilon_2 \pi_1} \qquad \text{Equation 3-6}$$

It is reasonable to believe that the cost of missing a breast cancer diagnosis in a patient who actually has the disease is higher than the cost of incorrectly diagnosing a patient as having the disease while actually s/he does not have it. The scenario is presented in a confusion matrix in Table 3-1

	Group	
Specialist's decision	1	2
1	No cost	C_2 (Lower Cost)
2	C_1 (Higher Cost)	No cost

Table 3-1: Misclassification costs

In other words, the value C in Figure 3-1 should be chosen such that either PM or CM is minimised. If Equation 3-3 is minimised, the resulting classification rule assumes equal costs of misclassification. Minimisation of Equation 3-6 will result in a classification rule that assumes unequal priors and unequal misclassification costs. It can be shown (Sharma, 1996) that the rule which minimises Equation 3-6 assigns an observation to Group 1 if

$$\frac{f_1(x)}{f_2(x)} \geq \left[\frac{C_2}{C_1}\right].\left[\frac{\pi_2}{\pi_1}\right] \qquad \text{Equation 3-7}$$

And to Group 2 if

$$\frac{f_1(x)}{f_2(x)} < \left[\frac{C_2}{C_1}\right].\left[\frac{\pi_2}{\pi_1}\right] \qquad \text{Equation 3-8}$$

This discussion provides a general framework for predictive modelling which can be

implemented using any of the predictive modelling methods in use. The present project uses three techniques; Logistic regression (LR), neural networks (NNs) and decision trees (DT).

In our predictive model we have used the above-mentioned techniques. Logistic regression is used in this model because the outcome of the project will be binary: the patient will either have breast cancer, or not. The aim of logistic regression is to predict the relationship between dependent and independent variables. In our case breast cancer is dependent and predictive factors are independent; this is suitable for our project as we need to predict whether patients have the disease or not.

Neural network is used when the relationship is more complicated or complex and is non-linear. In our case data are non-linear, and by implementing step function we will be able to get an output of either 0 or 1, meaning presence or absence of breast cancer for any patient.

The purpose of the decision tree is to break complex data into smaller subsets, which helps in identifying the presence or absence of any factor. DT can analyse homogenous and heterogeneous data. In our case it will help us in identifying or splitting patients into two groups: either they will have disease or not. From the literature it is evident that nobody has used these three techniques at the same time to compare the results and accuracy of the outcome. In our project we will be able to make conclusions which will be suitable techniques for predicting breast cancer. These techniques will be discussed in detail to achieve the second objective of the study ***(Develop predictive models appropriate for estimating risks of developing breast cancer).***

(Develop predictive models appropriate for estimating risks of developing breast cancer).

3.3 Logistic regression

LR is built on the foundations of linear regression, in which the aim is to predict the relationship between the dependent variable (Y) and a set of independent variables (X) as shown in Equation 3-9

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon \qquad \text{Equation 3-9}$$

Where β_0 and β_i are constants, ε is an error component and $X = (x_1 \ldots x_k)$ is the vector of inputs. LR, however, covers both linearity and non-linearity in determining the relationship between predictor variables (X) – which are usually continuous, categorical,

or both - and a dichotomously coded dependent variable (Y). The difference between linear regression and LR is that, while linear regression outcome is continuous, logistic regression outcome is binary. For example, if the logistic model outcome is denoted by Y and its input vector by X, the model is as in Equation 3-10.

$$\log \hat{} = Po + 2?=i\ PiXi \qquad \text{Equation 3-10}$$

Where p is the probability that dependent variable Y=1 and $Xt, / = 1,2,...,k$ are the independent variables (predictors) and the $p0,\ p2 \rightarrow Pk$ are the regression coefficients. Unlike ordinary linear regression, LR does not assume linearity. As graphically illustrated in Figure 3-2, it has an S-shape capturing linear, near-linear and non-linear scenarios.

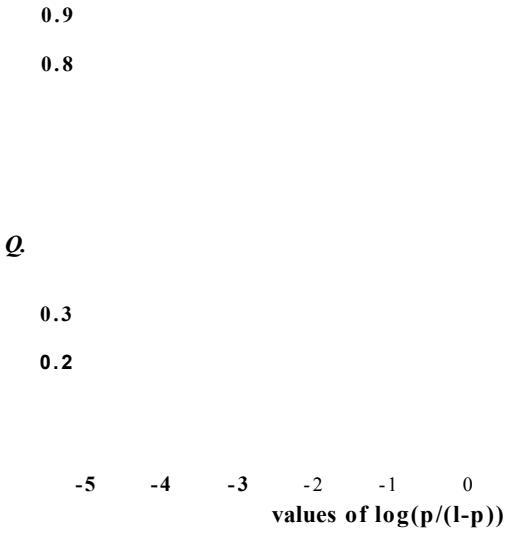


Figure 3-2: Logistic Regression functions

The procedure uses three main approaches to model fitting: forward, backward and stepwise regression. In the forward stepwise approach, variables are sequentially added to an "empty" model. In contrast, backward procedures start with all of the variables in the model, and proceed by eliminating the variables at each step. Only variables with significant effect will be retained in the final model. LR finds applications in a wide range of fields including the biomedical sciences, in which its use has increased in recent years. A typical application of the method in this field is the prediction of susceptibility to particular diseases. The present study uses the method in order to predict occurrences of breast cancer in the presence of geographical and ethnic factors.

The nature of LR makes it a natural choice for predicting the likelihood of an individual's developing breast cancer given information in the attributes (variables such as age, sex, family history, first birth, ethnicity and place of birth) in which two groups may be considered with rule allocating to group 1 with a probability greater or equal to 0.5 and to group 2 otherwise.

Some of the model's key parameters are the odds ratio (OR) estimate, the confidence interval (Cl) and the p-values. The OR estimates the odds of an event (such as having or not having breast cancer) occurring in one group to those of it occurring in another. Odds are a way of expressing the probability of an event in one group. In the present case, ORs can be used to measure the association between breast cancer and its predisposing factors as illustrated in Table 3-2

Risk factor		**Y**	
		Case(breast cancer)	Control(no breast cancer)
X	Did not smoke	a	b
	Smoked	c	d

Table 3-2: Inputs for the computation of odds ratios of patients with breast cancer based on whether they had ever smoked or not.

If **a** and **c** represent the respective conditional probabilities of "did not smoke" and "smoked" among those who have breast cancer and **b** and **d** the corresponding probabilities among those who are free of the disease, the odds of a patient with breast cancer being a smoker (***Odds.***j) and an individual without breast cancer being a smoker *(Oddsz)* can be defined as follows:

$$Odds_1 = \frac{c}{-} \quad and \quad odds_2 = \frac{d}{-} \qquad \text{Equation 3-11}$$

Consequently, the OR can be defined as follows:

$$OR \quad \_ \quad 0\text{Ł} \qquad \text{tion 3\_12}$$

$odds_2$

An OR equal to 1 means that the groups have equal probabilities of getting breast cancer, implying that smoking is not a predisposing factor, while an OR greater than 1 means that having smoked raises the chances of contracting the disease. On the other hand, an OR of less than 1 implies that smoking reduces the chances. The OR is a measure of association, which provides an insight into the underlying philosophy of LR, under which the model seeks to determine whether or not there are significant odds that an independent variable is associated with one of the groups.

Research studies Mechanic *et al.* (2007) and Jimmy *et al.* (2008) do not report ORs as precise numbers, but qualify them by giving CIs. These are ranges of numbers indicating the possible ranges of values that contain the true value with a given probability level, typically 95 per cent. Whether or not a given OR value indicates a significant association between the corresponding variables can be determined by a derived decision rule. In that sense, it is possible to generate a CI, as for the correlation coefficient. For instance, based on the entries in Equation 3-13 an association is significant when

$$|\log OR| > Z_{\alpha/2}\sqrt{1/a + 1/b + 1/c + 1/d} \quad \text{Equation 3-13}$$

Which is based on the $z-$criterion with a given α, e.g. $\alpha = .05$. In other words, it can be said that a proposed risk factor for the disease is significant risk if it generates an OR greater than 1 and the lower boundary of the CI does not go below 1.

Another crucial parameter is the p-value, which represents the probability that any observed difference between groups is due to chance. A p-value close to 0 indicates that the observed difference is unlikely to be due to chance, whereas the closer it gets to 1 the less the difference between the groups, and hence the greater the chance that any variation is random. Nichols et al. (2005) calculate ORs and 95 per cent CIs to evaluate the association of known risk factors and breast cancer among premenopausal women living in Vietnam and China using LR.

They observe an inverse trend between increasing parity and decreasing breast cancer risk (P = 0.002). Women ages ≥ 25 years at first birth had increased breast cancer risk compared with women ages <25 years at first birth (OR, 1.53; 95% CI, 1.20-1.95). Women who consumed alcohol had increased risk of breast cancer compared with women who did not (OR, 1.85; 95% CI, 1.32-2.61). They find that the distributions of age at menarche and body mass index are very different to those commonly observed in epidemiologic studies of Caucasian women. Also the distributions of these risk factors are consistent with the relatively low incidence rates of breast cancer in Vietnam and China as compared with the United States.

On the African continent such studies are still rare. Meanwhile, Okobia *et al.* (2006) evaluated the potential risk factors for breast cancer in Nigerian women using a case–control design of 250 women with breast cancer and their age-matched female controls. The association of risk factors with breast cancer was assessed using LR, the authors

State that positive family history of breast cancer in first- and second-degree relatives (Odds ratio [OR] = 8.07, 95% confidence interval [CI], 1.003, 64.95, p = 0.04), education of high school level and above (OR = 1.35, 95% CI 1.04, 1.74, p = 0.0205), age at first full term pregnancy greater than 20 years (OR = 1.32 95% CI 1.01, 1.71, p = 0.0413) and waist/hip ratio (WHR) (OR = 1.98, 95% CI 1.27, 3.10, p = 0.0026) were associated with increased risk of breast cancer. Their findings show that socio-demographic characteristics, reproductive variables and anthropometric measures are significant predictors of breast cancer risk in the target population.

The application of LR in the present study will therefore seek to consolidate breast cancer-related information from Africa by comparing it with similar studies across the world, especially in the developed world, where the disease has been more extensively studied. Neural networks, another model to be used in the present study, will now be discussed.

3.4 Neural networks (NNs)

NNs (NN) denote a mathematical device for modelling the relationship between a set of input variables (X) and an output (Y). They find application in almost every situation in which that relationship is complex and typically non-linear. They are also commonly referred to as artificial neural networks (ANNs) Ripley (1994) to distinguish them from their originator - biological NNs (Kononenko, 2001). The origins of NNs lie in the work of McCullough and Pitts (1943) who developed mathematical models based on the observational behaviour of biological neurons. Based on their findings, they started investigating whether and how physical systems could emulate the brain's neurons.

The relationship between ANNs and biological neurons is particularly interesting, as they are both constructed in a way that enables them to take in inputs and, based on a processing mechanism, trigger a response. A graphic illustration of the similarities between biological and artificial NNs is given in Figure 3-3, in which the synapses in the biological systems play the role of neuron signal junctions. In an ANN the synapses are equivalent to the input variable weightings, being represented by real numbers accounting for the importance of each input.

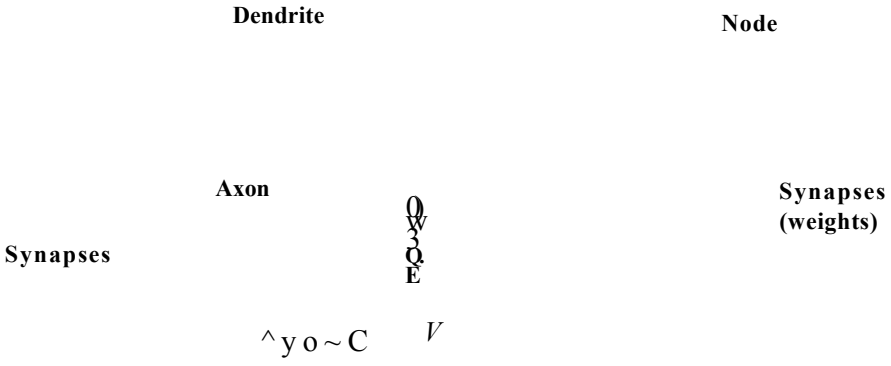


Figure 3-3: A comparative illustration of biological and artificial NNs (Ohno-Machadol996)

A typical neuron collects signals from others through a host of fine structures called dendrites. The equivalent of this in the ANN does not exist, as collection is manual. The neuron sends out spikes of electrical activity through a long, thin stand called the axon for the synapse to convert, but in the ANN model this does not happen as the synapse (i.e. the weighting) is attached to the node. NNs have evolved greatly since McCullough and Pitts (1943) originated the concept, as shown by Rosenblatt (1962), who developed the learning algorithm model and called it a simple perceptron.

This algorithm was based on McCulloch-Pitts model neurons with two layers, input and output. NNs suffered a major drawback following Minsky and Papert's (1969) highlighting of the limitations of the simple perceptron's linear separability. However, the idea survived, thanks to Hopfield (1982) who combined a number of previous NN ideas to form a mathematical analysis model having finite interconnected neurons with dual roles of both input and output. The popularity of NNs grew increasingly in the mid-1980s with the advent of the back-propagation algorithm by Rumelhart et al. (1986).

3.4.1 An overview of neural networks

The mechanics of ANNs can be understood by first describing some of the key concepts determining the network and relating to it:

- Nodes, interconnections and architecture: A node is a connection point typically linking input and output variables through different connections. NNs will usually have many nodes and interconnections that enable a complete design to be constructed, usually referred to as the network architecture.

have many nodes and interconnections that enable a complete design to be constructed, usually referred to as the network architecture.

- The mechanics of NNs take form when data are received through input nodes and are typically combined linearly with randomly initialised weightings by a combination function as illustrated in Figure 3-4.

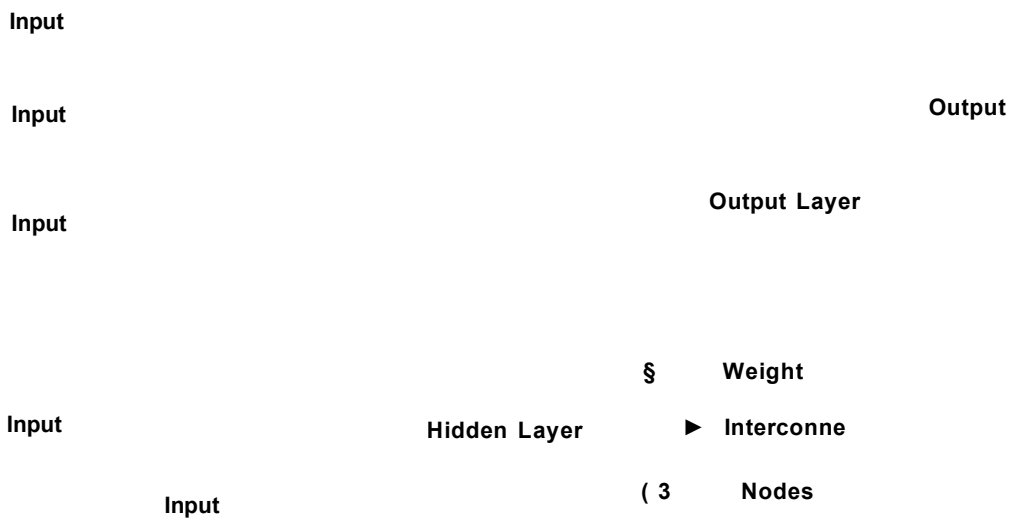


Figure 3-4: A graphic presentation of an artificial neural network

There are different types of NN in the literature typical examples of which include:

- The simple perceptron, a connection of nodes made up of linked input and output layers. Each connection has a weight that is adjustable when desired.

The multilayer perceptron includes an input layer, an unlimited number of hidden layers and an output layer. Jerez-Aragones *et al.* (2003) proposed a NN-based method for estimating the correct classification probability from a distribution using a multi-layer perceptron.

The output layer is usually a single node whose output is constrained within a preset limit based on the use of the activation function. The target variable (the output) is also well-defined, which means that what is being sought is clear. The number of nodes to

include in the hidden layer affects, to some degree, the accuracy of the results. For extremely complex networks with complicated interconnections, it is sometimes beneficial to use a large number of hidden nodes. There are no specific rules for determining how many nodes to include, but typically the most suitable hidden layer is determined by repeated testing. The model is graphically illustrated in Figure 3-5, which has one input layer, two hidden layers and one output. Removing the two hidden layers reduces the model to a simple perceptron.

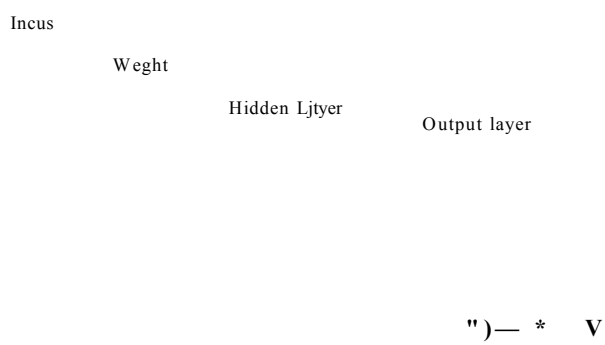


Figure 3-5: Multilayer ANN

Apart from the case of the simple perceptron, the randomly initialised weightings are usually modified to produce accurate predictions in what is referred to as the learning process, through which the NN model is shown data examples from which to learn its prediction rules. After the primary assignment of weightings, the output is compared to the known results and iterative training takes place. As the number of cases available for training increases, the degree to which the NNs can distinguish the various characteristics of the different inputs increases.

The inputs received by a single processing node can be represented as input vector X, where every element ^ represents the signal from one of the inputs. Weight is

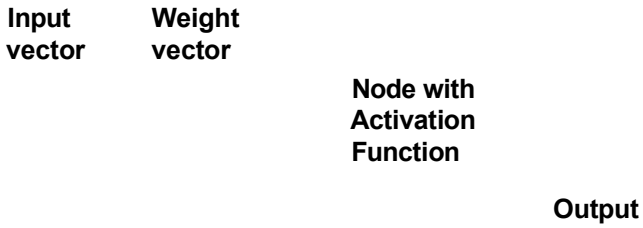


Figure 3-6: Simplified model of single processing node

The weighted sum is found as follows:

$$y = \text{Hi}\, x_i w_i \qquad \text{Equation 3-14}$$

The weighted sum in a multilayer NN is found as follows:

$$y = ZiXiWij \qquad \text{Equation 3-15}$$

Where w_{ij} in Figure 3-5 represents the weighting for the connection from an input or node i to nodey.In this case the inputs, for instance X_1, X_2, X_3, X_4 are variables such as age, family history, breast-feeding and place of birth. The usual ANN output is in the range of 0 to 1, where 0 and 1 represent the absence or presence of breast cancer respectively. Any output value above this range indicates the presence of an abnormality, while a value under it indicates the absence of any abnormality.

There are three typical activation functions: step and saturation functions and hyperbolic tangent. Each adjusts the output in a specific way, with the largest difference occurring in the step function.

- The step function takes the weighted input and applies the following conditions:

$$/(*) = \{1 \; ifd.iXiWi > 0), 0 \; if(ZiXiWi < 0)\} \qquad \text{Equation 3-16}$$

The function in Figure 3-7 pushes all incoming signals to either a 0 or a 1. The advantage of using this function is that all outputs are defined as either positive or negative. This function can make data that are not properly trained produce output classed as positive or negative.

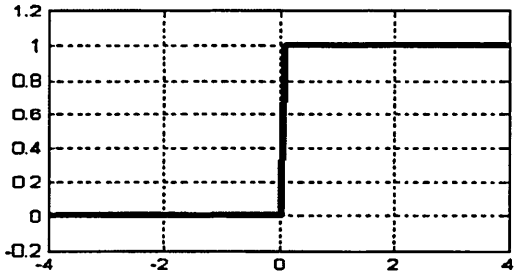


Figure 3-7: Step function

- The saturation function takes the weighted input and applies the following condition:

$$f(x) = \frac{1}{1+e^{-Bx}} \quad \text{Equation 3-17}$$

Where **β** determines the slope and is determined by the ANN as seen in Figure 3-8.

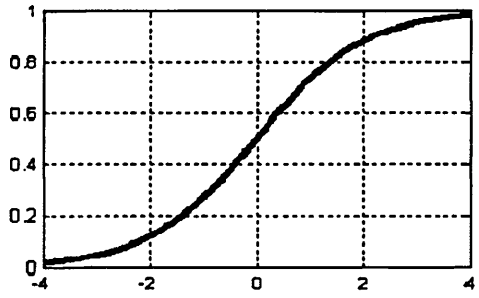


Figure 3-8: Saturation function

- The Hyperbolic Tangent Function takes the weighted input and applies the following condition:

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \text{Equation 3-18}$$

This function can be seen in Figure 3-9.

Figure 3-9 : Hyperbolic tangent function

All the activation functions have the same purpose, receiving the information from the input layer and manipulating it into a specific range. The saturation function limits the output to between **0** and **1**, while the hyperbolic tangent function gives an output range of -1 to 1. On the other hand, the step function, does not output a range, but only one of the values **0** or **1**.

- Radial basis function networks typically have three layers: an input layer, a hidden layer with a non-linear radial basis function activation function and a linear output layer. The value of this function depends only on the distance from the origin; the rule is typically taken to be the Euclidean distance and the basis function is understood as Gaussian.

- Back-propagation is a general method of teaching ANNs. It consists of the propagation of errors starting at the output layer, through the hidden layer, and so on to the input layer, in a backward direction. The training process continually minimises the error possibility by adjusting the weightings, as shown in Equation 3-19.

$$W_{new} = w_{old} + \eta(d - y)x \qquad \text{Equation 3-19}$$

where **d** is the desired output (the target), y is the calculated output, $(d - y)$ is the error GO, w_{new} is the new updated weighting, w_{old} is the previous weighting and x is the input to the processing element. η is the leaming-rate parameter; the leaming-rate is a positive constant limited to the range $0 < \eta < 1$

Changes in weighting are associated with interconnections, and result in the known data being processed again. ANNs continue to refine the weightings until the errors produced from processing the known dataset are minimised, at which point the training process ends. In this manner NNs avoid over-training.

3.4.2 Neural Networks in biological and medical research

The use of ANNs in biological and medical research has increased tremendously in the last few years. In the area of medicine, ANN research includes:

- Detection of medical problems. ANNs have been used to recognise predictive patterns so that appropriate treatments can be prescribed.
- The predictive and pattern recognition abilities of ANNs make them naturally suitable for use in medicine. They have found applications in clinical diagnostics Kordylewski *et al* (2001), image analysis and interpretation Ozkan *et al* (1993) and even drug development Grosan *et al* (2006).
- Prediction of diagnoses, including several types of cancer Maclin *et al* (1991) and Wilding *et al* (1994) as well as diagnoses of appendicitis, back pain, dementia, psychiatric disorders, acute pulmonary embolism and temporal arteries (Wei *et al* , 1998).
- Prognoses, such as valve-related complications in heart disease (Wilson *et al*, 1998).
- Determination of risk or disease profiles (Kannel *et al* 1996).
- NNs are an ideal candidate for adaptation into a breast cancer detection scheme as a pre-processor to the more time-intensive imaging algorithms. A database of known cases is required for training the NN, after which unknown data can be provided for diagnostics.
- It is necessary to estimate the risk of relapse for breast cancer patients Furberg *et al.* (2002), since it affects the choice of treatment. This problem involves analysing the time taken for a patient to relapse, relating this to predictive variables. It is possible to predict the risk that a patient will relapse within a certain amount of time, although the exact time when such a relapse occurs cannot be determined.

3.5 Decision Tree

The predictive problem highlighted in [3.2] can also be tackled using the decision tree method, a collective term used to describe Classification and Regression Trees (CART). Both types of tree learn the rules from the training data and use them to carry out predictions. Classification arises in the case of category-dependent variable and predicts class membership of new cases, while the latter carries out similar predictions on the basis of a continuous dependent variable. In both cases, the model sequentially splits the training dataset into a number of supersets based on selected data variable thresholds (typically one at a time), the selection of which depends on the adopted measure of

impurity (see below). Its name derives from the fact that the model assumes a tree structure growing from root to leaves (Breiman, 1984).

A decision tree model typically starts by breaking the whole dataset into two groups which are subsequently broken down into smaller sub-groups, each being at least as pure as the data group from which it derives. To split the data, the model needs some knowledge of what each of the classes is like. This information is obtained through the learning process, by which data at any level are analysed to find the independent variable that most distinguishes the classes. An illustration of the tree partitioning process is provided using the synthetic cancer risk factor matrix in Table 3-3, in which the age ranges from, say, 18 to 75 and body weight ranges from, say, 45kg to 110kg. Then, if the data matrix is denoted by R, the decision tree prediction of breast cancer can be illustrated as in Figure 3-10 in which there are only two classes, sick and healthy.

Age	Weight	Breast Feeding	Menache
A1	W1	B1	M1
A2	W2	B2	M2
A3	W3	B3	M3
An	Wn	Bn	Mn

Table 3-3: An illustration of breast cancer risk factors

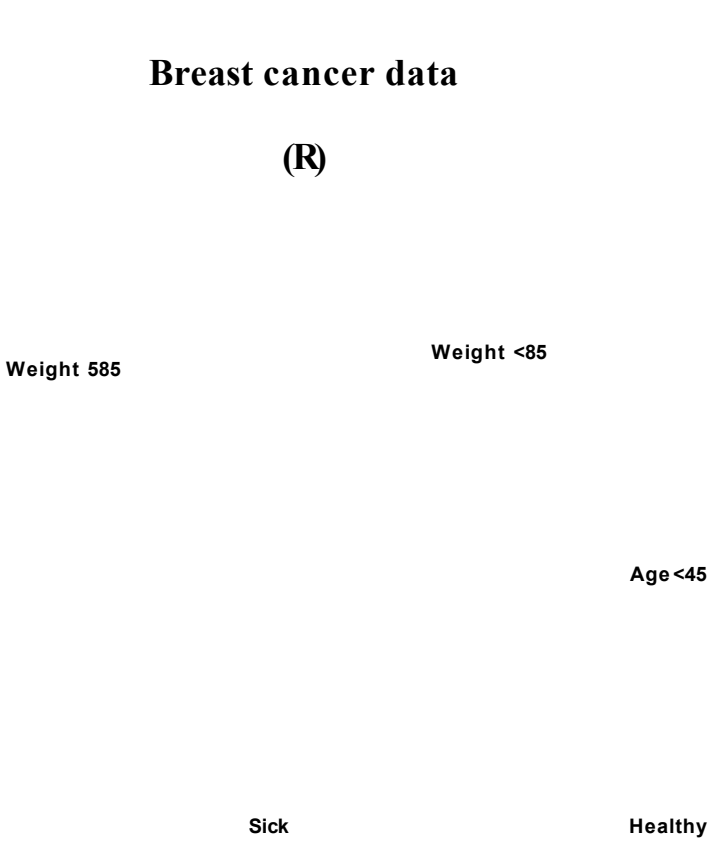


Figure 3-10: Decision tree: a simple structure

Each circle in the tree in Figure 3-10 represents a node. A decision tree grows from the root node (1) which contains all the instances in the training set and goes on to split the data at each level to form new nodes. In this case the nodes 2, 4 and 5 are called terminal nodes, or leaves, as they not split any further. Terminal nodes play a special role when the tree is used for prediction, as they collectively determine the model's accuracy and reliability.

Each node contains information about the number of instances at the node and the distribution of the dependent variables. As shown in Figure 3-10, the tree model initially splits the training dataset on the basis of the variable Weight using the rule "Allocate to sick class if the patient's weight is greater than or equal to 85 and to healthy class if the weight is less than 85". The split yields one terminal node (2), and another (3) which is split further based on the rule "Allocate to sick class if the age of the patient is greater than or equal to 45 or to healthy class if the age is less than 45", which yields the terminal nodes 4 and 5. As noted earlier, splitting is based on variable thresholds - taken one at a time - and an adopted measure of impurity, which will now be discussed exposition.

3.5.1 Measures of impurity

Decision tree mechanics are determined by measures of impurity. The impurity of a set of samples is designed to capture how similar the samples are to each other and to try to minimise the variance over a partitioning of the data multiple each part with the number of samples will encourage larger partitions, which the present research finds leads to better decision trees in general. There are several ways to measure degrees of impurity, the most common being Gini and Entropy, as briefly described below.

The Gini coefficient was developed by Italian statistician Corrado Gini in 1912, as described by Hu *et al.* (2008), to measure how often any given element from a dataset would be incorrectly labelled if that labelling were to be randomly apportioned according to the distribution of labels in the subset. In other words, it measures the probability that two individuals chosen at random belong to the same group. It is computed by multiplying the probability of each item being chosen by the probability of a mistake in categorising that particular item. The measure reaches its maximum value when group sizes at the node are equal; the Gini index equals zero when all cases in the node belong to the same class.

The Gini Impurity Index can be computed as in Equation 3-20

$$G = \sum_{i=1}^{k} p_i(1 - p_i) = \sum_{i=1}^{k} p_i - \sum_{i=1}^{k} p_i^2 = 1 - \sum_{i=1}^{k} p_i^2 \quad \text{Equation 3-20}$$

where p_i is the groups proportion and k is the number of groups.

As regards in Figure 3-10 in which k =2 if 100 patients are divided into two groups, with and without breast cancer, according to the weight variable with the divider at greater than or equal to 85 group have the disease or free from the disease if the weight is less than 85, the result of splitting the root node is that node (2) consists of 38 patients with the disease and two without it, while node (3) consists of 58 patients without the disease and two with it.

Node (3) is split further, based on the age variable, so that if the age is greater than or equal to 45 have the disease or free from the disease, the result of splitting node (4) is that there are 20 patients with the disease; node (5) consists of 36 patients without the disease and four with it. In this case the Gini Index for node (4) is equals zero because all cases in this node belong to the same group, which means that the node is pure and cannot be split further. The Gini Index for node (5) not equals zero that is because it is

node is not pure and may split further on the basis of another variable, thus the best Gini splits try to produce pure nodes.

The Entropy is another measure of the impurity introduced by Shannon (1948), as cited by Golan (2002). It is related to Information Theory, in the sense that the higher the level of entropy, or uncertainty, of some data, the more information is required in order to completely describe that data. In other words, it measures the impurity of a node and the homogeneity of the dataset. The entropy of a particular decision tree node is the sum, over all the classes represented at that node, of the proportion of records belonging to a particular class multiplied by the base two logarithm of that proportion. This sum is usually multiplied by -1 to obtain a positive number. The entropy can be computed according to Equation 3-21.

$$E = -\sum_{i=1}^{k} p_i \log_2 p_i \qquad \text{Equation 3-21}$$

A node with higher entropy than another is more heterogeneous and therefore less pure. Entropy is zero if all members of the collection belong to the same class. If the collection contains unequal numbers, the entropy is between zero and one, and it reaches maximum value when all classes have equal probability.

Entropy for node (4), mentioned above, is zero, which means that the node is pure and will not split further. In the decision tree the aim is always to decrease the entropy of the dataset until leaf nodes are reached, at which point the remaining subset is pure - in other words, it has zero entropy and represents instances of one class only.

Each of the above-listed criteria may be used to split a tree. The splitting process may continue until each observation is individually classified - which is not a desirable outcome as, although the training error may go down to zero, the model cannot be generalised to new observations as the results are data-specific, as illustrated in Figure 3-11. To avoid this condition, a decision tree model is controlled either by imposing growth restrictions or by growing a maximum size tree and then pruning it down to size using a number of techniques such as cross-validation.

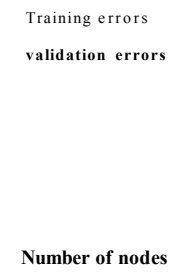


Figure 3-11: Determining when over-fitting begins

3.5.2 Tree pruning and model generalization

The situation in Figure 3-11 is described as data over-fitting which, as noted above, may be avoided by controlling the tree growth. Over-fitting means the data set only recognises the training instances, and never learns to classify new instances. It can easily lead to predictions that are far removed from the range of the training data.

There are various ways to avoid over-fitting in decision trees, including:

- Stopping tree growth before it reaches the point where it perfectly classifies the training data (pre - pmning).
- Allowing the tree to over-fit the data, and then post-pruning it.

Algorithms that build trees to maximum depth will automatically raise pruning.

Pre- pruning determines when to stop the growth of a tree, while post-pruning reduces the size of a fully expanded one. Although the former approach might seem more direct, post-pruning has been found to be more successful in practice due to the difficulty of estimating precisely when to stop growing the tree. One other way of avoiding over-fitting is to validate the model during its training, which can be achieved by a process known as cross-validation. This process estimates the expected level of fit of a model to a data set different to the one used to train the model.

3.6 A proposed strategy for data analysis

This section describes the strategy used to analyse the breast cancer data described above. It consists of two main steps in the first of which all data attributes are divided into demographical, geographical, controlled, uncontrolled and health condition categories. The controlled attributes refer to those variables that can be controlled by humans, such as weight, types of vegetables and meat, sporting activities, whether and for how long infants are breastfed, number of children, age at last pregnancy and duration of oral contraceptive use (where used), while the uncontrolled attributes are those that cannot be controlled. They include such aspects as height, work involving contact with radiation, age at menarche, age at first pregnancy, age at menopause and spontaneous abortion. Health condition is a three-level variable indicating whether or not the patient has other diseases, a family history of breast cancer, or a family history of other genetic conditions. The second step involves testing the models for their relevance in predicting breast cancer.

A graphic illustration of the two steps is shown in Figure 3-12, on which each of the three models will be tested with respect to the data attributes' relevance to predicting breast cancer.The significant variables are used in predicting breast cancer each time, testing for significance. The model is re-tested with the significant variables while the insignificant ones drop out until the final model is obtained.

Furthermore a new step model will built called hybrid model, taking the best factors from the decision tree model output and passing them into the neural network architecture in order to produce a classification result. A second hybrid model uses the same factors identified from the decision tree stage and passes them to a logistic regression model, in which the results are easier to interpret structurally than in the neural network. Finally, the results of the base models (the decision tree, single neural network and logistic regression models) will be compared with the hybrid model results.

The cross-validation method, which randomly divides the data into K-groups using one subset for validation and the remaining K-1 subsets for training the model, was applied to test the model. In this case, cross-validation was used to compare the performances of different folds in order to determine power of the model in predicting breast cancer. Typically, the model with the lowest generalised error was selected.

More specifically, trees of different sizes were built for the decision tree method. The test error was measured for each tree, and only the tree with the minimum error rate was considered. The significant risk factors for decision trees are only those involved in constructing trees using deviance value[9]. This technique was applied to the plan depicted in Figure 3-12.

The forward NN method was thought most applicable for selecting the most important factors. In this procedure, a NN model was firstly trained on a single factor. Each risk factor's contribution was evaluated by its error rate. All the possible models, each of which was based on one variable (i.e. one risk factor), using the error rate, was tested. The factor showing a minimum error will be included in the next models, which will consist of two factors. This process continues until no further reduction in error rate is noted.

3.7 Data description

Data consisting of 3,057 data samples on 29 variables were obtained from four cancer treatment centres in Libya during the period 2004-2008, as shown in Figure 3-13 of which 1,563 constitute patients diagnosed with breast cancer. The four cancer centres – Sabratha, Benghazi, Tripoli and Misurata – lie on the Mediterranean and they care for patients coming from as far south as Gat near the Algerian border and Al Qufra near the Sudanese and Chadian borders. Thus, the data were sampled from the Libyan population of around six million people scattered over 1.8 million square kilometres exhibited a steady increase in the rate of breast cancer over the period of time, with the highest growth occurring in Benghazi and the lowest in Misurata.

Data has been collectected by visiting these centres. These centres holds the record of patinets being treated in centre, access was provide to this data from the specialist after explanation of aims and objectives of the study and how the result will help them in

9 This term describes the fitness of statistics for a model that is often used for statistical hypothesis testing

making good plan for fighting the disease, confidinality of the data was assured to relevent personels, if any thing was missing author of this thesises posted the questionerie to the patinet or contacted the if there is any missing information in folder of the hospital. Detail of data collection is in Appendix B and evidence of data collection and questionnaire is given in appendix C.

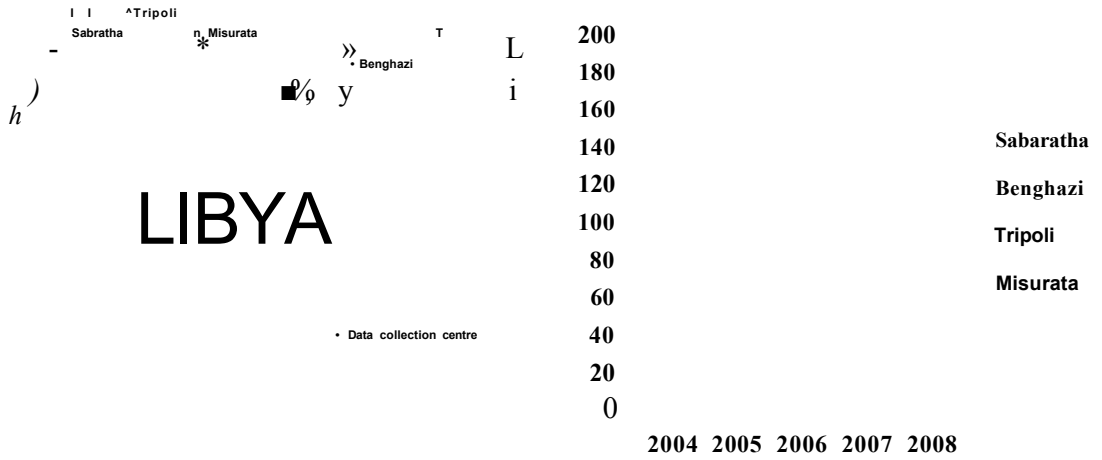


Figure 3-13: The general trends of breast cancer in Libya (constructed from sampled data)

The sampled data from the four centres are generally placed in a matrix of the form exhibited in Equation 3-22 for domain-partitioning purposes. To uncover breast cancer patterns, we apply a combination of classification tree models to learn the rules from the training data and use them to carry out predictions. The models sequentially split the training dataset X into groups based on selected data variable thresholds. To split the data, the model needs some knowledge of each of the classes.

$$X = \begin{matrix} Xn.xn.xiv. \\ X2VX22'X23' \quad .X_{2p} \\ \\ XN\backslash'XN2'XN3 \quad XNp \end{matrix} \qquad \text{Equation 3-22}$$

3.7.1 Data from the African oncology institute - Sabratha

Figure 3-14 shows the number of patients suffering from various types of cancer who had been treated in the Sabratha institute during 2004. Of 497 patients in this category, 67 (13.48 per cent) had breast cancer. However, Figure 3-15 shows that breast, lung and prostate cancer and Hodgkin disease are the most common types of cancer and that breast cancer came second after lung. The figure also shows that breast cancer occurred

most often in those aged 35-44 years.

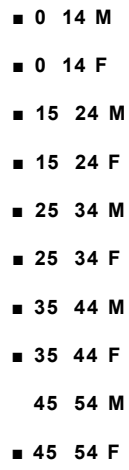


Figure 3-14: All cases of cancer by age in 2004 at Sabratha Institute

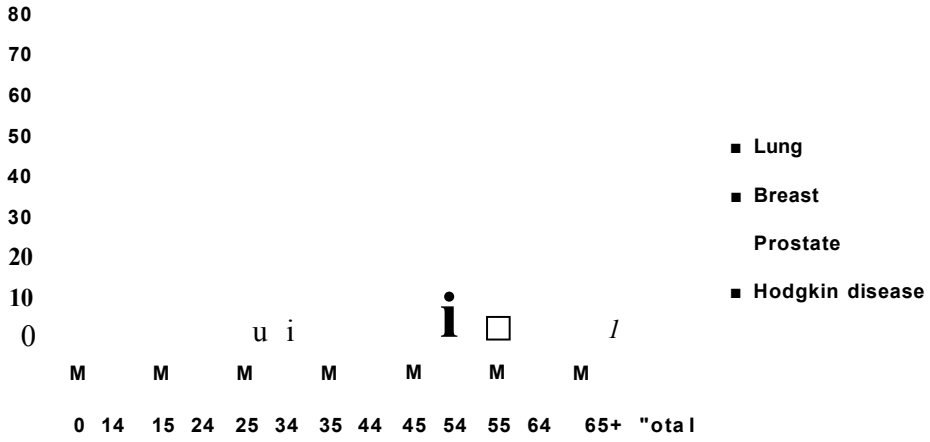


Figure 3-15: Four most prevalent types of cancer by age group in 2004 at Sabratha Institute

In 2005 the number of breast cancer patients was 80 (12.98 per cent), out of 616 of all types of cancer (Figure 3-16). Figure 3-17 shows that cancer of the colon overtook the other four, and that breast cancer came third after colon and lung cancer respectively. Breast cancer was also most prevalent that year among those aged 45-54.

Figure 3-16: AH cases of cancer by age in 2005 at Sabratha Institute

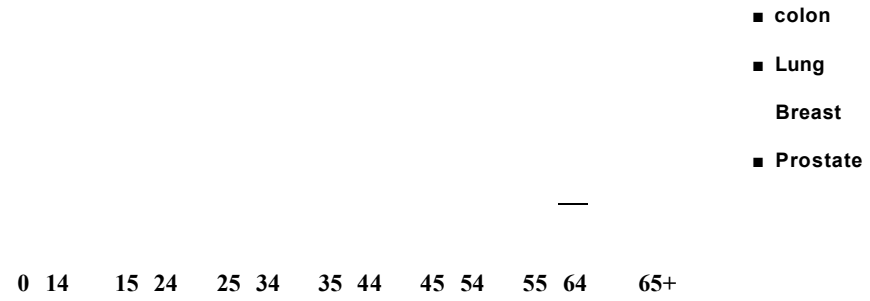


Figure 3-17: Four most prevalent types of cancer by age group in 2005 at Sabratha Institute

In 2006 leukaemia became one of the four most prevalent types of cancer (Figure 3-18, but again breast cancer accounted for the greatest number of cases - 101 (20.15 per cent) out of 501 cancer patients. Figure 3-19 shows that the age group most prone to breast cancer was the 35-44 year old one.

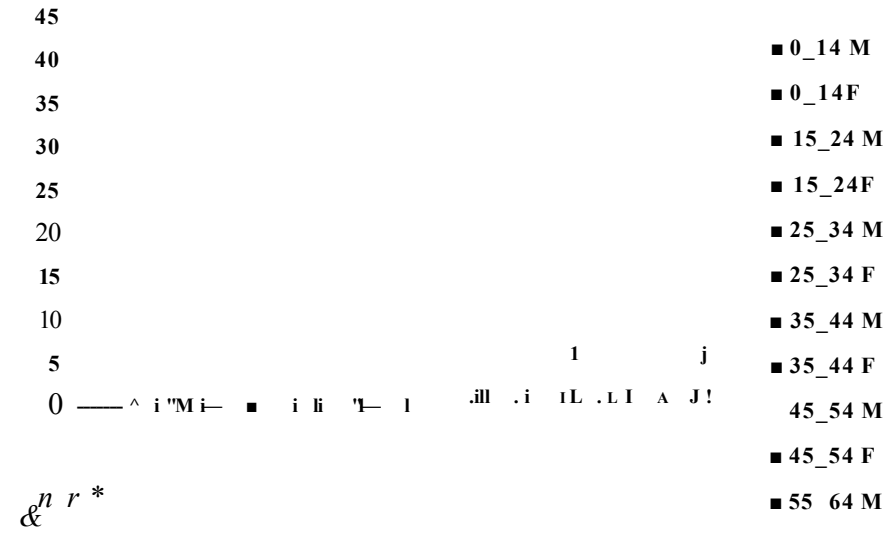


Figure 3-18: All cases of cancer by age in 2006 at Sabratha Institute

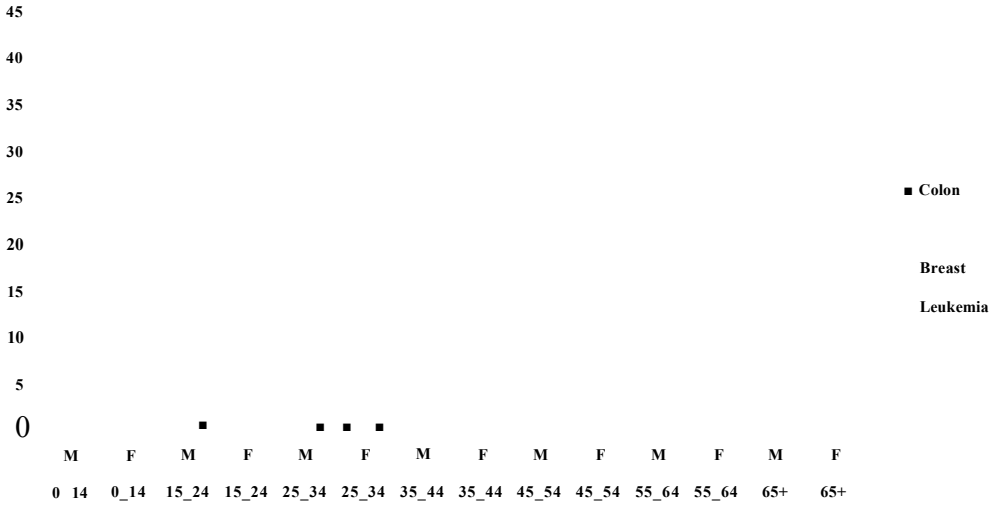


Figure 3-19: Four most prevalent types of cancer by age group in 2006 at Sabratha Institute

During 2007 breast cancer remained the most prevalent type, with 105 (17.5 per cent) of 600 patients having contracted this form (Figure 3-20). Figure 3-21 shows that breast cancer came fourth after cancers of the colon, rectum and lung, and were followed by prostate cancer; the most susceptible age group was 35-44 years.

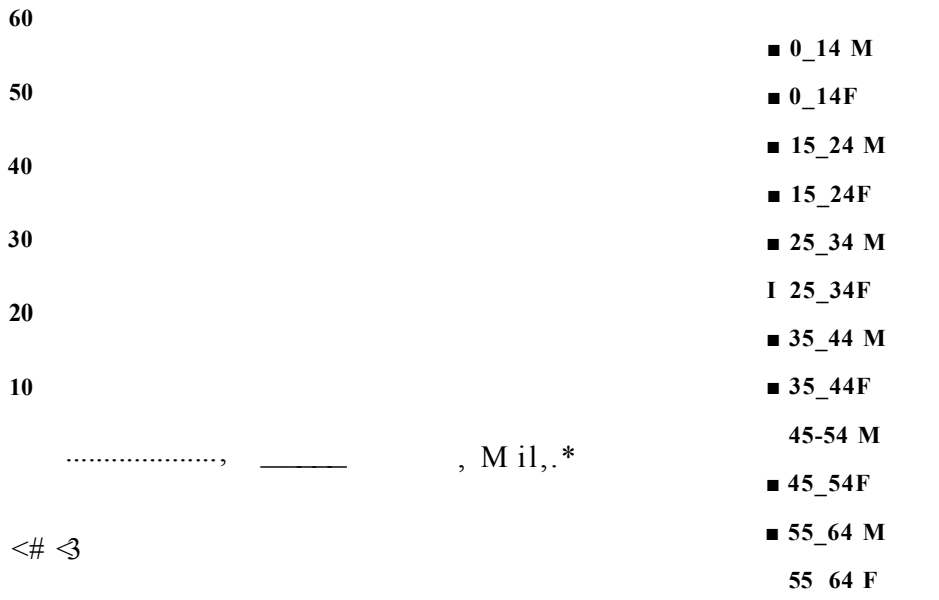


Figure 3-20: All cases of cancer by age in 2007 in Sabratha

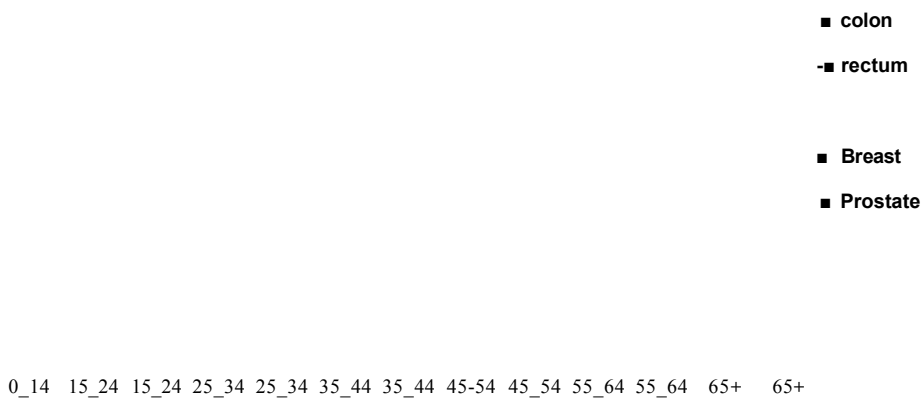


Figure 3-21: Five most prevalent types of cancer by age group in 2007 at Sabratha Institute

Breast cancer patients numbered 162 (26.51 per cent) of the total 611 treated in 2008 (Figure 3-22). Breast cancer came third after cancer of the colon and lung cancer, and was followed by prostate cancer. Figure 3-23 shows that the most common age group for breast cancer patients was 45-54 years.

Figure 3-22: All cases of cancer by age in 2008 at Sabratha Institute

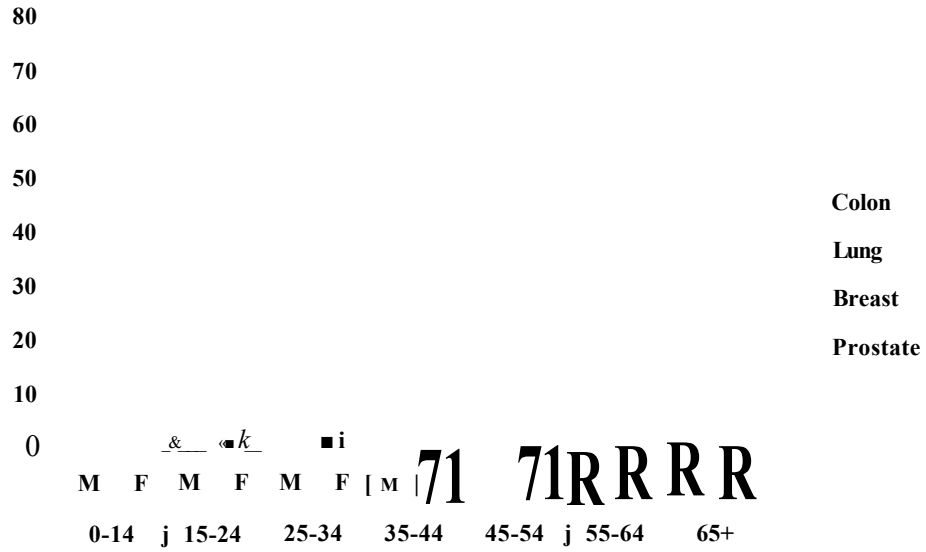


Figure 3-23: Four most prevalent types of cancer by age group in 2008 at Sabratha Institute

3.7.2 Data from the Benghazi centre

There were 95 breast cancer patients in 2004 at the Benghazi centre, constituting 23 per cent of the total 410 cancer patients (Figure 3-24). Breast cancer cases the most prevalent, followed by cancers of the lung, colon, rectum and ovary. Figure 3-25 shows that breast cancer patients aged 50-59 years old were the most susceptible to the disease.

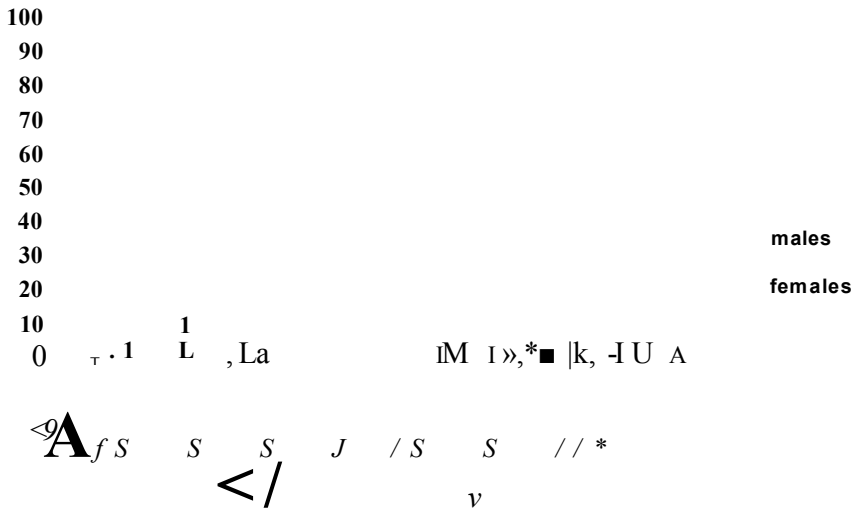


Figure 3-24: All cases of cancer by gender in 2004 at Benghazi centre

Figure 3-25: Cases of breast cancer by age group in 2004 at Benghazi centre

In 2005 there were 112 breast cancer patients, constituting 22.71 per cent of the total 493 cancer patients (Figure 3-26). Breast cancer was the most prevalent type, followed by cancers of the lung, colon and prostate; the most susceptible age group was 40-49 (Figure 3-27).

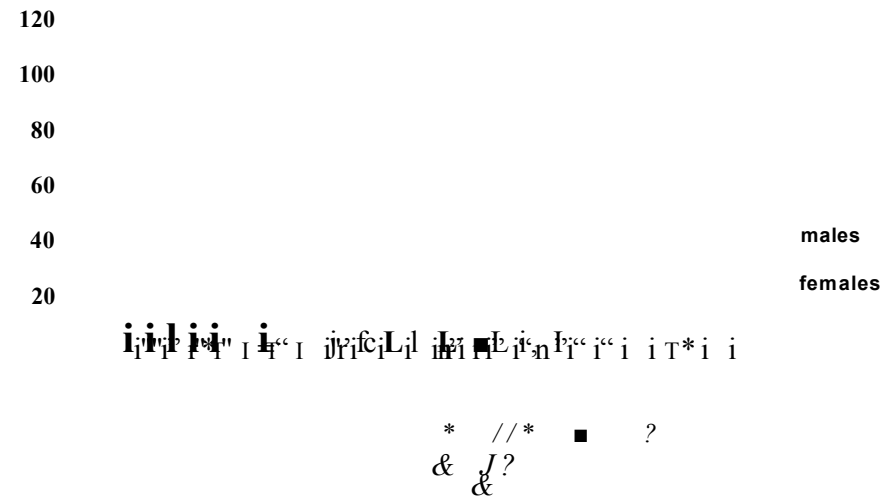


Figure 3-26: All cases of cancer by gender in 2005 at Benghazi centre

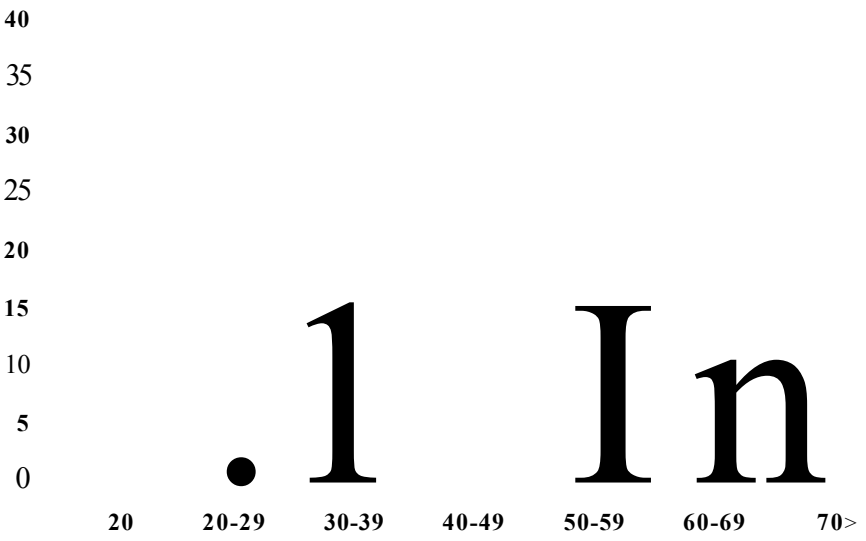


Figure 3-27: Cases of breast cancer by age group in 2005 at Benghazi centre

During 2006 breast cancer remained the most prevalent type, with 113 (19.85 per cent) of the total 569 cancer patients (Figure 3-28). The main age group of sufferers was 40-49 years (Figure 3-29), and this form of the disease was still the most prevalent, followed by cancers of the lung, colon and prostate.

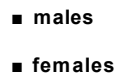


Figure 3-28: All cases of cancer by gender in 2006 at Benghazi centre

Figure 3-29: Cases of breast cancer by age group in 2006 at Benghazi centre

The total number of breast cancer patients in the year 2007 was 148, 20.96 per cent of the total 706 cancer patients (Figure 3-30). Breast cancer again accounted for the greatest number of patients, with those aged 40-49 years being the most highly represented (Figure 3-31).

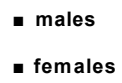


Figure 3-30: All cases of cancer by gender in 2007 at Benghazi centre

Figure 3-31: Cases of breast cancer by age group in 2007 at Benghazi centre

The total number of patients suffering from breast cancer in the year 2008 was 173, 24.53 per cent of the total 705 cancer patients (Figure 3-32). Breast cancer was again the most prevalent, followed by cancers of the lung, colon and prostate. Figure 3-33 also shows that the mean age of breast cancer patients was 50-59.

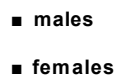


Figure 3-32: All cases of cancer by gender in 2008 at Benghazi centre

Figure 3-33: Cases of breast cancer by age group in 2008 at Benghazi centre

3.7.3 Data from the central hospital in Tripoli

There were 308 patients suffering from breast cancer disease during the period between 2004 and 2008; the yearly totals are shown in Figure 3-34. Figure 3-35 represents the number of cases by age group. The breakdown results in proportions of breast cancer patients of 1.62 per cent of those aged under 20, 8.11 per cent of those aged 20-30, 26.62 per cent of those aged 31-40, 33.11 of those aged 41-50, 29.2 per cent of those aged 51-60 and 1.29 per cent of those age over 60.

Figure 3-34: Number of breast cancer patients by year at Tripoli centre

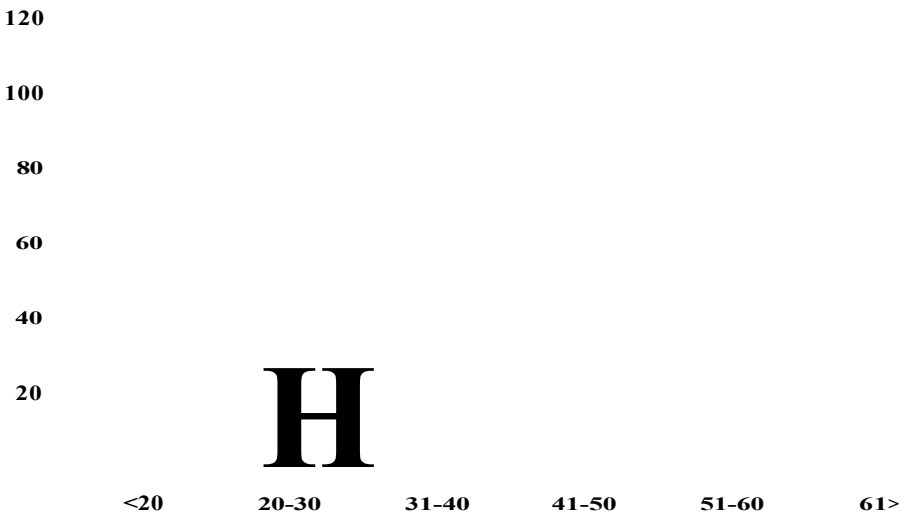


Figure 3-35: the number of breast cancer Cases by age group at Tripoli centre

3.7.4 Data from the National Cancer Institute in Misurata

There were 136 patients suffering from breast cancer in Misurata institute during the period between 2004 and 2008 (Figure 3-36). It can be deduced from Figure 3-37 that 2.20 per cent of the patients aged less than 20, 8.82 per cent of those aged 20-30, 30.14 per cent of those aged 31-40, 47.05 per cent of those aged 41-50, 8.08 per cent of those aged 51-60 and 3.67 per cent of those aged over >60 suffer from breast cancer.

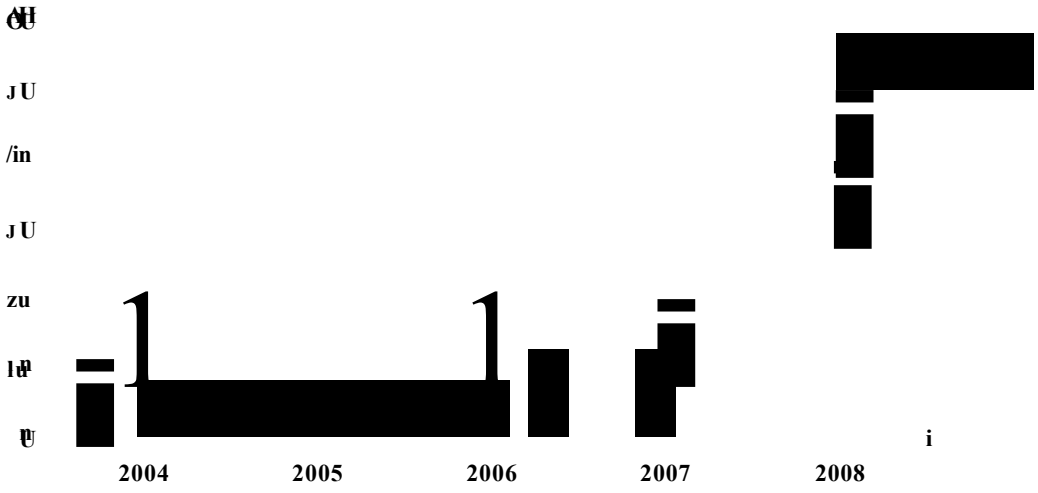


Figure 3-36: Number of breast cancer patients by year at Misurata centre

Figure 3-37: the number of breast cancer Cases by age group at Misurata centre

In general the data shows that women younger than 20 years old run the least risk of developing breast cancer, while those aged 35-54 are at the greatest risk. As we mentioned in Chapter **three** (3.6), the **data will be divided into five group** (demographical, geographical, controlled, uncontrolled and health condition) to test the performance of the three data techniques.

3.8 Selection of R software Program

Many statistical programs are available for analysing the breast cancer data including SAS, MATLAB, SPSS, EXCEL, etc. The main software used in this study is **R** supported by EXCEL. The R software provides a wide variety of statistical and graphical techniques; linear and nonlinear mixed effects models, statistical tests, clustering and classification. In addition, it allows users to add additional functionality by defining new functions, and easy to integrate it into existing functionality.

3.9 Research ethics

Data used during this research was collected from four cancer centres in Libya. Prior to collection, extensive meetings were held with heads of departments, to whom the purpose of research and its benefits for the Libyan community and more generally for the African continent was explained. The written permission for the collection and use of data shown in the Appendix was granted. This research uses highly confidential and private data. The researcher assured the authorities that he will not misuse or reveal that data to any person not involved in this research. Personal information from all patient files was made available. During and after this study the researcher also followed the ethical code of the UK's Sheffield Hallam University.

Bryman (2008) emphasises that ethical considerations are critical and appropriate for any research process. Ethical issues were taken into account throughout all the stages of research, particularly during the data collection phase. Black (2002) notes that ethical considerations must generally be considered during the research design. Each person involved in the research is also considered to have certain roles and responsibilities.

While researchers should maintain high standards to ensure that data is accurate, and should not misrepresent that data, they are also required to protect the right to confidentiality of participants in the research (Zikmund, 1999).It follows that researchers' primary ethical consideration is to protect participating organisations and individuals from any possible disadvantages or adverse consequences that may result from their research (Black, 2002; Zikmund, 1999; Bryman, 2008).

3.10 Summary

This chapter has given detailed discussion about the three methods used in this research and their use in predicting breast cancer. Strategy along with data analysis is given in

this chapter, which also explains how data were collected during this research. The proposed strategy for data analysis which would apply is given in Chapter 4, and Chapter 5 has further analysis relating to the hybrid model.

Chapter 4 : Data Analysis

4.1 Introduction

This chapter provides data analysis of the sampled data from two perspectives. Firstly, we carry out an exploratory analysis of the data from the four cancer centres to obtain general patterns of the disease in Libya. Secondly, the sampled data are subjected to domain-partitioning using the methods and strategy outlined in Chapter three.

Our devised Strategy has been discussed in detail in Chapter 3, and the methodology in Section 3.6. The same strategy is followed in the analysis of the data. Data from 3,057 patients in various centres throughout the Libya, for cancer generally and breast cancer in particular, was analysed. Cross-validation was used to verify the results of this data using logistic regression, neural networking and decision trees. The results of the three-, five- and ten-fold cross-validation algorithms are discussed below. The dataset of breast cancer was divided into three, five and ten-folds in order to compute the error rate for the three classification methods using an algorithm of forward variable selection. For three-fold cross-validation the 3,057 patients were divided into three subsamples, one of which acted as a validation set containing data from 1,019 patients while the data from the remaining 2,038 in the other two was treated as training data. For five-fold validation, 611 patients were used for validation with the remainder constituting training samples. Finally, for ten-fold validation, the size of the subsample was reduced to 305 patients. The results of the logistic regression verification will be presented first, then those for neural networks and decision trees.

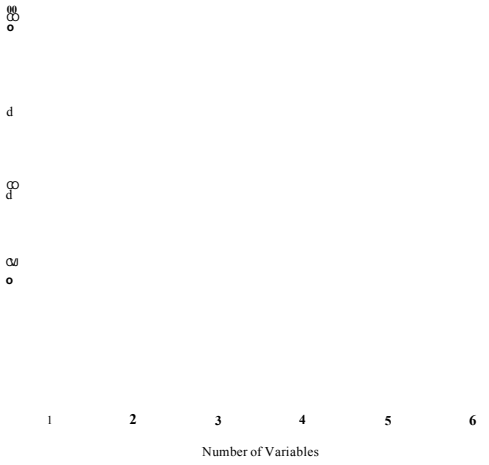
4.2 Logistic Regression

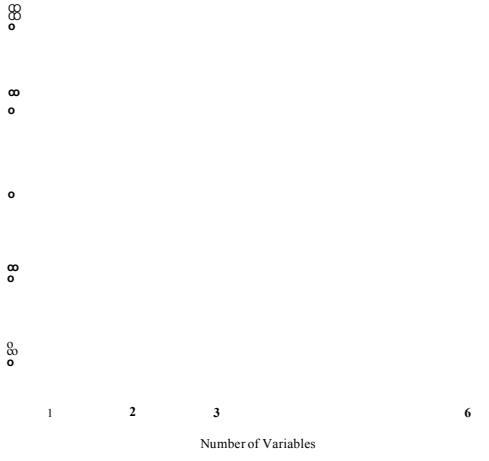
In this section, the results are organised according to the plan of selection algorithm. Each table shows the validation error evaluated by three, five and ten folders.

To determine the best number of variables leading to the lowest error rates, the logistic classifier is trained on all the possible subsets using three-, five- and ten-fold cross-validation. The experiment is repeated (bootstrapped) 50 times using the selection strategy, after which the odds ratio is estimated for the final classifier based on the best variables belonging to the factors under consideration.

4.2.1 Demographic factors

According to Fig 4-1, all the cross-validation folds display a very slow variation in error rate when the logistic classifier includes one or two variables. The variation becomes larger when three or more predictors are entered into the model, despite the error rate decreasing. The logistic classifier performs best when the classifier retains five or sex predictors. It should be noted that the error rate variation produced by ten-fold cross-validation seems to be somewhat lower that the results from three and five-fold. One reason for this may be attributed to the large number of training sets, which enables the logistic classifier to receive more information, particularly when the features of dataset are highly correlated, leading to poor parameter estimates.





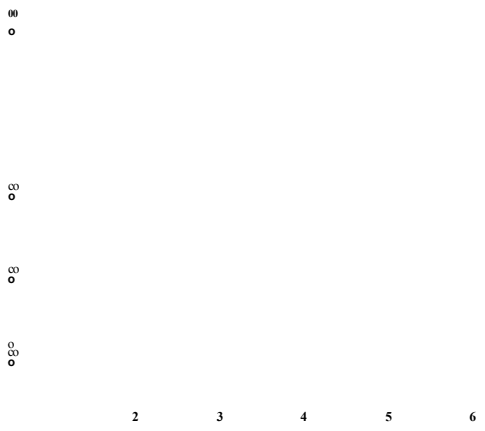


Figure 4-1: Error rates based on the number of variable for logistic classifier for three folds (top plot), five folds (middle plot) and ten folds (bottom plot) in terms of demographic factors.

Based on the selection results of demographic factor given in Table 4-1 the socio-economic variable plays the greatest role in distinguishing between the two groups for a simple model consisting of one variable; however the performance of this model is only mediocre. By adding the other important variables using the selection algorithm, the selection process is terminated at the socio-economic, educational level, age and employment variables, since no significant reduction in error rate is achieved by any further additions. In fact, error rates obviously do not drop markedly by adding the variables in the successive models. The number of folds evaluating error rate show a very similar level of performance.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Socio Economic	1	0.386	0.386	0.386
Education, level	2	0.355	0.355	0.355
Age	3	0.346	0.346	0.344
Employee	4	0.319	0.321	0.317
Marital State	5	0.305	0.307	0.303
Gender	6	0.301	0.303	0.298

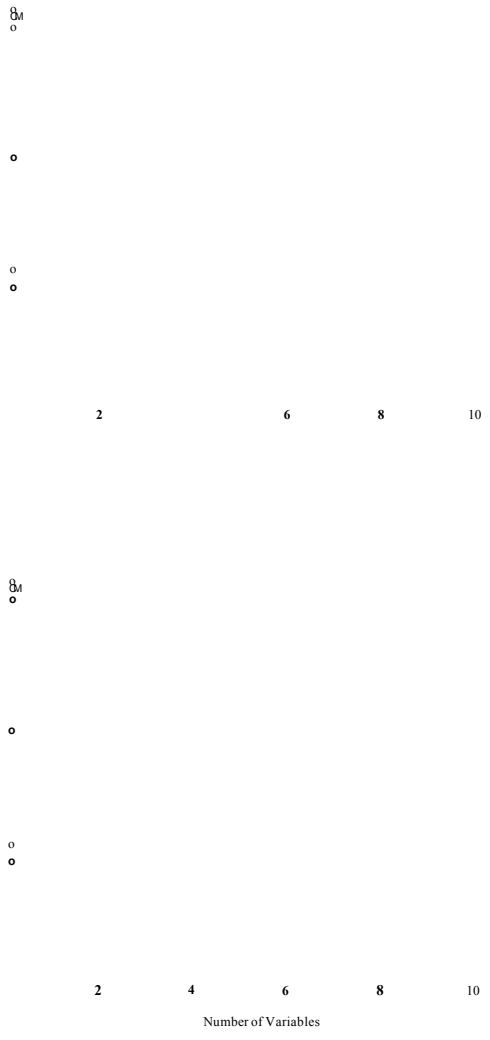
Table 4-1: Error rate of cross validation according to the logistic models trained on models demographic factor.

4.2.2 Controlled factors

The results of the controlled factors shown in Fig 4-2 demonstrate that the test error rates resulting from the K- (K=3, 5 and 10)-fold cross-validations follow the same pattern. The Interesting result, here, is that when complexity of logistic classifier, the reduction in error rate becomes obviously slow. The variation in error rate for selected subsets of predictors is quite low, resulting in stability of estimated parameters for all

bootstrapped samples.

According to Table 4-2, weight has the highest performance for simple models (about 79 per cent) - a considerable accuracy rate for a model based on one independent variable. Hence, weight is one of major predictive factors for the risk of developing breast cancer. The best logistic regression models based on two and three variables, breastfeeding and length of breast feeding are the most important variables respectively. The error rate for the model based on three variables drops to 13.7 per cent (about 7.2 per cent better than the simple model based on weight). By adding the other important variables using the selection algorithm, the selection process is terminated at the five variables of weight, breastfeeding, length of breastfeeding, sporting activity and kinds of meat, because the reduction in error rate for the other variables is relatively low.



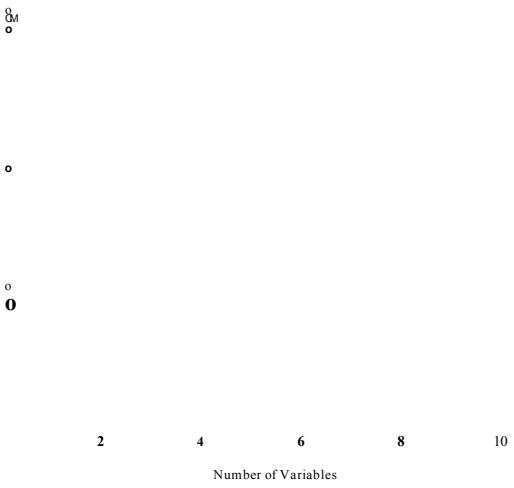


Figure 4-2: Error rates based on the number of variable for logistic classifier for three folds (top plot), five folds (middle plot)) and ten folds (bottom plot) in terms of control factor.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Breastfeeding	2	0.169	0.169	0.169
Length of Breastfeeding	3	0.137	0.137	0.137
Sport	4	0.110	0.110	0.110
Kind of Meat	5	0.085	0.085	0.085
Age at last pregnancy	6	0.079	0.077	0.078
Duration of oral contraceptive.use	7	0.077	0.074	0.073
Kind of Vegetable	8	0.074	0.069	0.065
Number of children	9	0.069	0.067	0.064
Oral contraceptive use	10	0.069	0.066	0.063

Table 4-2: Error rate of cross validation according to the logistic models trained on controlled factors

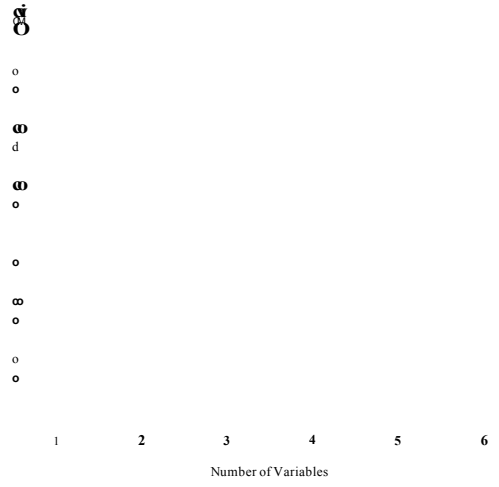
4.2.3 Uncontrolled factor

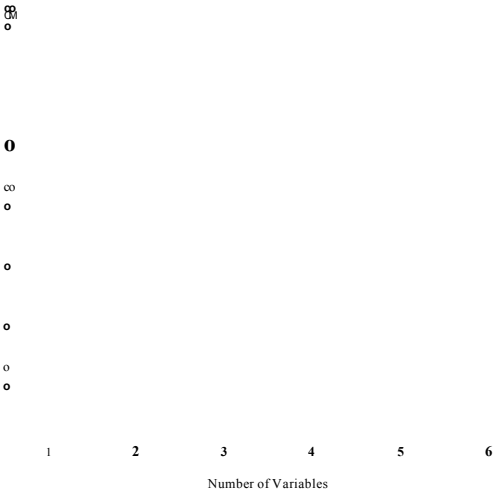
According to Fig 4-3 the variation in error rate is very low for all samples of cross-validation meaning and selected subsets of predictors. As long as the selected number exceeds three, the selection process does not increase the performance of logistic classifiers. However, the predictors of uncontrolled factors are not correlated with each other, as indicated by the low variation in error rates. Table (4-3) shows that height is the most important distinguishing variable, followed by work connected with radiation and then miscarriages. The error rate drops from 20.7 per cent for the simple model to 9.2 per cent for the one based on the three variables. The selection procedure was

stopped at this point, since the error rate does not change thereafter.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Height	1	0.207	0.207	0.207
Work connected with radiation	2	0.113	0.113	0.113
Spontaneous abortions	3	0.092	0.092	0.092
Age at menarche	4	0.092	0.092	0.092
Age at menopause	6	0.091	0.092	0.092
Age at firs Pregnancy	7	0.091	0.091	0.091

Table 4-3: Error rate of cross validation according to the logistic models trained on uncontrolled factors





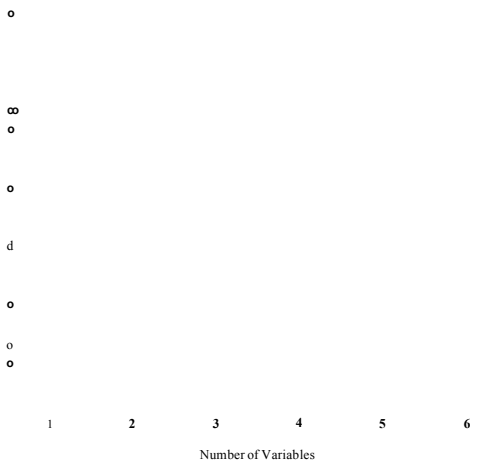


Figure 4-3: Error rates based on the number of variable for logistic classifier for three folds (top plot), five folds (middle plot)) and ten folds (bottom plot) in terms of uncontrolled factor.

4.2.4 Health factor

Fig 4-4 demonstrates no benefit from adding more than one predictor to the logistic classifier. All the bootstrapped samples in the K-fold cross-validations result in the same value of error rate. This is a very rare result.

The results in Table 4-4 show that, for a simple model, family history of breast cancer is the most important variable, and is thus largely responsible for development of the disease, with a prediction accuracy of 82.2 per cent. Other diseases and inherited disease variables do not contribute to the classification performance.

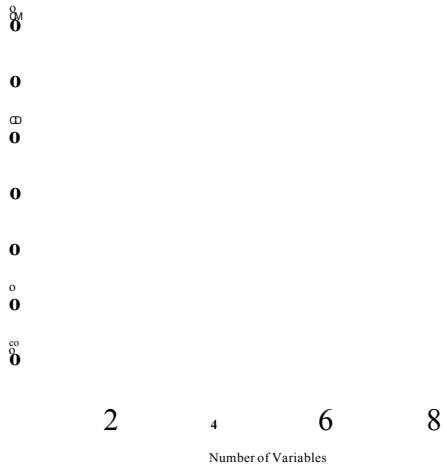
Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	**1**	**0.178**	**0.178**	**0.178**
Other diseases	**2**	**0.178**	**0.178**	**0.178**
Inherited diseases	**3**	**0.178**	**0.178**	**0.178**

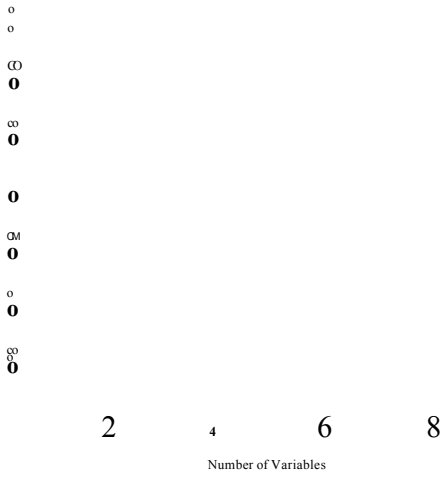
Table 4-4: Error rate of cross validation according to the logistic models trained on health factors

Figure 4-4: Error rates based on the number of variable for logistic classifier for three folds (top plot), five folds (middle plot)) and ten folds (bottom plot) in terms of health factor

4.2.5 Demographic and control factors

The results in Fig 4-5 show that error rate variation is reduced when using five- or 10-fold cross-validation, leading to stability in the values of the estimated parameters. For the cases of cross-validation, low variation levels will rise when there are more than four predictors.





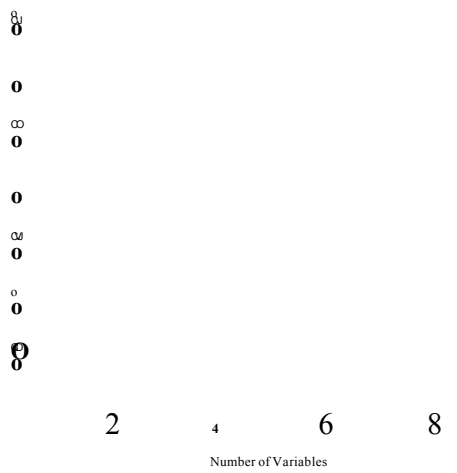


Figure 4-5: Error rates based on the number of variable for logistic classifier for three folds (top plot), five folds (middle plot)) and ten folds (bottom plot) in terms of demographic and control factors

This analysis is used to select the important variables from the retained ones of demographic and controlled factors. Table 4-5 shows that weight is the first variable entered into the model, and its error rate is 20.9 per cent using three, five and ten folds. The error rate drops significantly after entering breastfeeding, to 16.9 per cent, and decreases gradually when the other important variables are added using the selection algorithm. The selection process is terminated at the five variables of weight, breastfeeding, length of breastfeeding, sporting activity and kinds of meat, as shown in Table 4-5.The reduction in error rate demonstrates by these variables is relatively low; all variables provided under the algorithm of demographic selection are excluded from the final selection. In other words, the best controlled factors will be selected for final model. According to this result, the effect of the demographic factors has probably passed through the controlled factors.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Breastfeeding	2	0.169	0.169	0.169
Length of Breastfeeding	3	0.137	0.137	0.137
Sport	4	0.110	0.110	0.110
Kind of Meat	5	0.085	0.085	0.085
Socio Economic	6	0.078	0.078	0.078
Age	7	0.076	0.076	0.077
Employee	8	0.075	0.075	0.075
Education level	9	0.073	0.074	0.073

Table 4-5:Error rate of cross validation according to the logistic models trained on best demographic and control factors

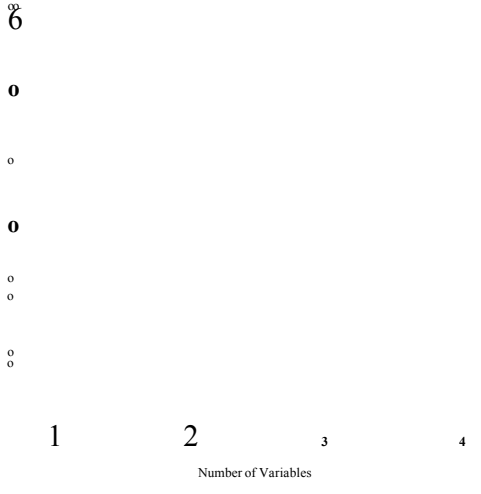
4.2.6 Uncontrolled and health factors

Similarly to the previous results, Fig 4-6 confirms that the larger the size of training set, the lower the level of variation. Hence, good estimation is gained, Three-fold cross-validation does provide a good result, although less efficient than 10-fold. Table 4-6 demonstrates the results of variables selection algorithm that provides the best uncontrolled and health factors for the construction of a model. Family history comes first, with an error rate of 17.8 per cent according to the three types of fold. This is consequently considered as a major predictive factor for breast cancer.

Family history is combined with height in a second model, where the error rate becomes 8.4 per cent- a good reduction. Adding work connected with radiation further reduces the error rate, although not as much. Because spontaneous abortions do not significantly improve the cancer classification, the selection process is terminated at the three variables of family history, height and work connected with radiation, as shown in Table 4-6 the reduction in error rate resulting from these variables is relatively low according to the three types of cross-validation.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Height	2	0.084	0.084	0.084
Work connected with radiation	3	0.069	0.069	0.069
Spontaneous abortions	4	0.068	0.069	0.068

Table 4-6: Error rate of cross validation according to the logistic models trained on best uncontrolled and health factors



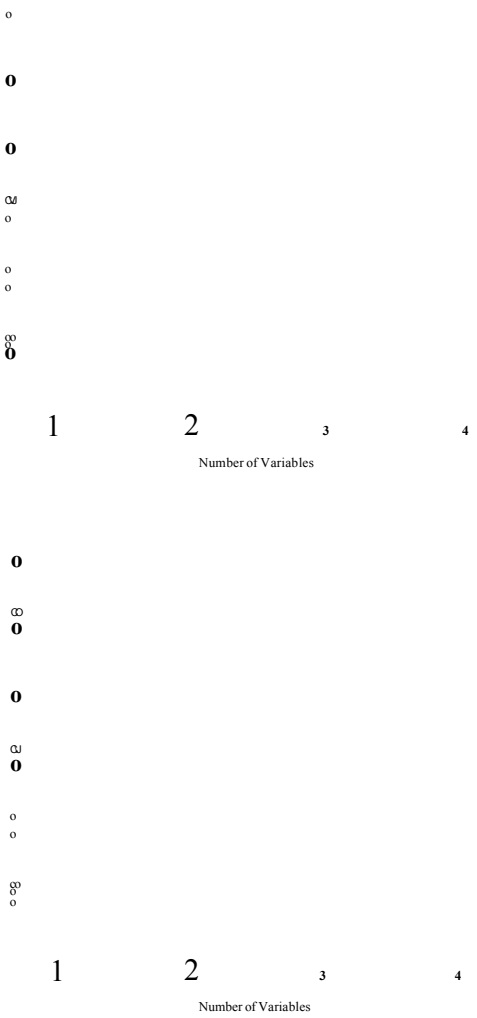
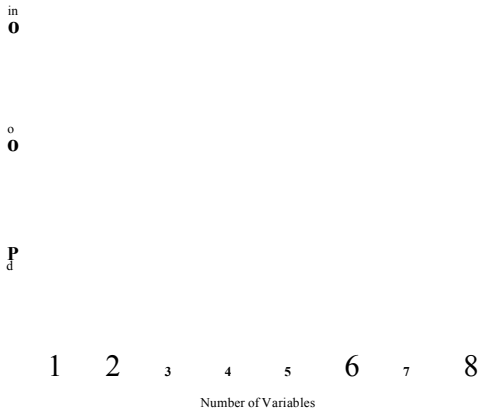
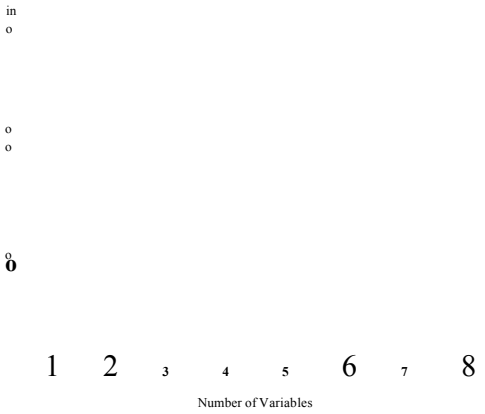


Figure 4-6: Error rates based on the number of variable for logistic classifier for three folds (top plot), five

4.2.7 All Factors

For the final model, Fig 4-7 leaves no doubt that 10-fold cross-validation does not result in a markedly more consistent estimation of parameters than does three and five-fold cross-validation when the number of predictors entered in the logistic classifier increases. Overall, K (where K=3, 5 and 10)-fold cross-validation using the bootstrap technique provides a good indicator for: assessing the consistency of estimated parameters and determining the number of predictors required for the best performance.





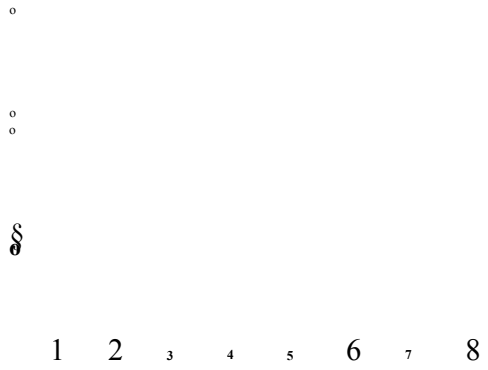


Figure 4-7: Error rates based on the number of variable for logistic classifier for three folds (top plot), five folds (middle plot)) and ten folds (bottom plot) in terms of all factors.

The best variables obtained from Table 4-5 and 4-6 are used to form a new combination in order increase the classification performance. Table 4-7 shows that the uncontrolled and controlled factors make the highest contribution to the classification. Family history and height result in error rates of 8.4 per cent. Length of breastfeeding from the uncontrolled factors, together with the variables already selected for this model, reduce the error rate to 6.9 per cent. In the writer judgment, the best performance of 96.27 per cent is achieved by adding work connected with radiation and breastfeeding to the previous variables.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.171
Height	2	0.084	0.084	0.084
Length of Breastfeeding	3	0.069	0.069	0.069
Work connected with radiation	4	0.046	0.046	0.046
Breastfeeding	5	0.037	0.037	0.037
Kind of Meat	6	0.034	0.034	0.034
Sport	7	0.027	0.027	0.023
Weight	8	0.021	0.021	0.021

Table 4-7: Error rate of cross validation according to the logistic models based trained on the all important factors

The potential for developing the disease will now be illustrated by the results of logistic regression. All the significant variables kept in the final model are used to produce the estimates of the logistic model. Each predictive variable is encoded into a number of binary variables taking the values of zero and one. If the original predictor is based on category *k,* the *k-1* binary variable is obtained. This procedure enables one to predict the

chance that this disease is present for each category using the odds ratio. Applying maximum likelihood to fit this model, Table 4-8 shows the results.

Variable	Category	Coefficient	Odds	SD-Error	Z-value
Weight in kg	<60	0	1		
	60-80	0.9740	2.6484	0.8420	1.157
	80-100	3.5216	33.8369	0.8698	4.049***
	>100	17.3955	358716	683.9363	0.025
Sport	Yes	0	1		
	No	2.9013	18.1971	0.3286	8.829 ***
Length of breastfeeding	No	0	1		
	< 1 year	-9.3952	0.0008	1.9415	-4.839***
	1 year	-10.4036	0.0003	1.9185	-5.423***
	≤ 2 year	-13.4009	0.0000	1.9934	-6.723***
Kind of meat	Red meat	0	1		
	Fish	-1.7170	0.1795	0.4343	3.953***
	Bird	-2.5597	0.0773	0.3912	-6.542***
Breastfeeding	Yes	0	1		
	No	6.4641	6.4148	1.1998	5.387***
Height	≤150	0	1		
	151-170	3.0286	20.669	0.4418	6.855***
	>171	6.6959	8.0905	0.5758	11.628 ***
Family history	No	0	1		
	Yes	3.9778	0.01872	0.3680	10.808 ***
Work connected with radiation	No	0	1		
	Yes	3.8076	0.0220	0.4276	8.905 ***

Significant keys: * significant at 0.05, ** significant at 0.01 and *** significant at 0.001

Table 4-8: Results from the final model of logistic regression fitting breast cancer dataset.

This table includes coefficient, standard error and odds ratio as well as Z score (coefficient divided by standard error) for the coefficient in the model. This score is used to test the null hypothesis that the coefficient in the model is zero. The table shows women who weigh between 60-80 kgs to have an almost two and a half times greater risk of developing the disease. The odds of the weight group of 80-100 kg is 33.84, which means that the potential for developing breast cancer is about 33.84 times the likelihood of it not occurring. The results for women who weigh more than 100 kg are surprising, and must be explained with some care. While not a significant number, this

group has the highest odds ratio for developing the disease, but it is important to note that as weight increases, so does the chance of developing breast cancer. This agrees with the results of Adebamowo et al. 2003, Stephenson and Rose (2003), Sweeney et al. (2004) and Krebs et al., (2006).

Women who do not practise sport are three times more likely to contract the disease; this agrees with the results of Gilliland et al (2001) and John et al. (2003). However, it differs from the results of Colditz et al. (2003), Margolis et al. (2005) and Mertens et al. (2006), who found no relationship between physical activity and breast cancer in pre- or post-menopausal women.

A diet that includes fowl has a lower odds ratio than one with other kinds of meat, thus we verify the results of Zografos et al. (2004), and Hanf and Gonder (2005). Those who do not breastfeed their babies are at six times the risk of those who do, agreeing with the results of Zheng et al. (2000) and the Collaborative Group on Hormonal Factors and Breast Cancer (2002), whereas the chance of contracting breast cancer diminishes markedly as the length of breastfeeding increases. The odds of infection are about four times higher if there is a family history of breast cancer; these results agree with Sattin et al. (1985), Pharoah et al. (1997), Calderon et al. (2000), the Collaborative Group on Hormonal Factors in Breast Cancer (2001) and Ebrahimi et al. (2002), who declared that breast cancer risk was significantly greater in women with a family history of the disease. Also, the odds relating to work connected with radiation are about four times higher; these results agree with those of Furberg et al. (2002), who declared that no matter how low the radiation dose , there is some small risk associated with exposure.

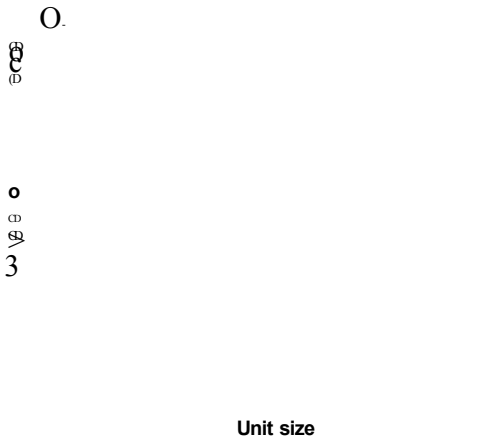
4.3 Neural Network

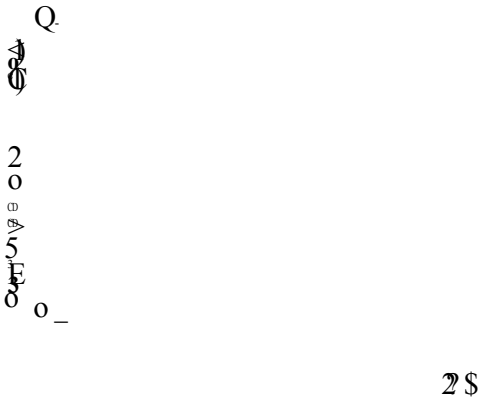
Before starting to apply the selection procedure, the number of hidden units minimizing the error rate using a cumulative error rate difference will be determined. This strategy can enable us to reduce the time to compute the results. The selection procedure will be based on the number of hidden units discovered.

4.3.1 Demographic factors

Figure 4-8 shows that the difference for all folds for hidden units of size two can result in low cumulative error rates. Notice that the performance of the neural network will be somewhat low if the size of units becomes large, say seventeen. The results of all the folds for variable selection algorithms demonstrate that the classifier trained on the

socioeconomic variable results in the highest performance. By constructing new classifiers, each consisting of two variables, one of which is the socioeconomic one, the best reduction in error rate is obtained by the classifier trained on the socioeconomic and education level variables, as shown in Table 4-9. By following the same process the socioeconomic variable and the level of education are selected. The variables of employment and age are retained in the next step of analysis, since they show error rates of 26.4 per cent. The performance of the excluded factors of marital state and gender is not noteworthy.





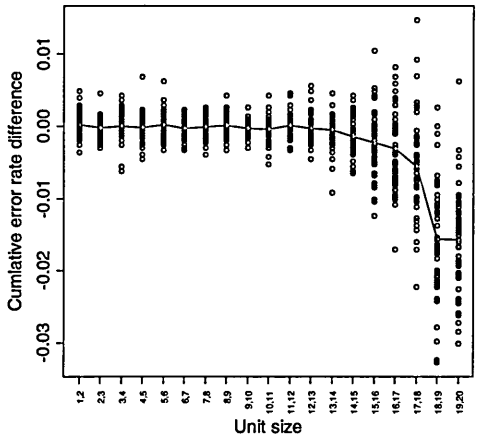


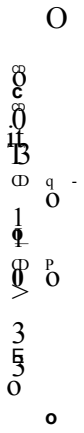
Figure 4-8: cumulative error rates based on the number of hidden units for three folds (top plot), five folds (middle plot) and ten folds (bottom plot) in terms of demographic factors.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Socio Economic	1	0.386	0.386	0.386
Education, level	2	0.355	0.355	0.355
Employee	3	0.310	0.311	0.313
Age	4	0.264	0.269	0.265
Marital State	5	0.252	0.258	0.248
Gender	6	0.246	0.258	0.248

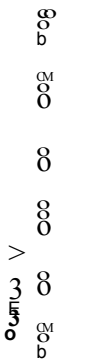
Table 4-9 Error rate of cross validation according to the neural network models trained on demographic factors

4.3.2 Controlled Factors

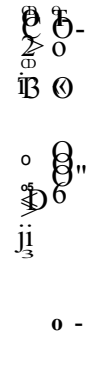
The best size for hidden units is two (Fig 4-9). Variable selection shows that the classifier trained on the weight variable leads to the highest performance among the all simple classification models of neural network. By constructing new classifiers, each of which consists of two variables, one of which is weight, it is apparent that the best reduction in error rate is obtained by the classifier trained on weight and breastfeeding as shown in the Table 4-10. After the selection algorithm was run several times, weight, breastfeeding, length of breastfeeding, sporting activity and kinds of meat were selected as variables for further analysis, since they show an error rate of 8.5 per cent. The excluded factors are the number of children, Age at last pregnancy, kinds of vegetable and use of oral contraceptives.



Unit size



Unit size



Unit size

Figure 4-9: cumulative error rates based on the number of hidden units for three folds (top plot), five folds (middle plot) and ten folds (bottom plot) in terms of control factors.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Breastfeeding	2	0.169	0.169	0.169
Length of Breastfeeding	3	0.138	0.139	0.139
Sport	4	0.113	0.112	0.111
Kind of Meat	5	0.085	0.086	0.087
Kind of Vegetable	6	0.082	0.082	0.082
Duration of oral contraceptive.use	7	0.078	0.080	0.082
Number of children	8	0.077	0.079	0.071
Age at last pregnancy	9	0.075	0.076	0.076
Oral contraceptive use	10	0.075	0.075	0.075

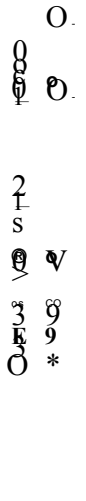
Table 4-10 Error rate of cross validation according to the neural network trained on controlled factor

4.3.3 Health Factors

Two hidden units are enough to reach a good performance, as shown in Fig 4-10. The selection of classifier based on a single variable demonstrates that the health factor of family history performs better than the other variables; its error rate is 17.8 per cent. The construction of new classifiers trained on two variables, one of which is family history, shows that the best reduction in error rate is obtained by the classifier trained on the family history and other diseases variables, as shown in Table 4-11 in. Family history, other diseases and inherited disease are retained as factors for further analysis.

Unit size

Unit size



Unit size

Figure 4-10 cumulative error rates based on the number of hidden units for three folds (top plot), five folds (middle plot) and ten folds (bottom plot) in terms of health factors

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	**1**	**0.178**	**0.178**	**0.178**
Other diseases	**2**	**0.143**	**0.143**	**0.143**
Inherited diseases	**3**	**0.132**	**0.132**	**0.132**

Table 4-11: Error rate of cross validation according to the neural network model trained on health factors.

4.3.4 Uncontrolled factors

Fig. 4-11 shows that two hidden units can be used for constructing the classifier. The height variable provides the highest performance among the simple classifiers. The construction of new classifiers consisting of two variables, one of which is height, demonstrates that the best reduction in error rate is achieved by the classifier trained on height and working with radiation (Table 5-12). Following the same procedure, height, working with radiation and spontaneous abortions are the factors chosen for further analysis, their error rate is 9.61 per cent. Age at menarche, age at first pregnancy and age at menpause do not show remarkable performances, so they are excluded from the uncontrolled factors model.

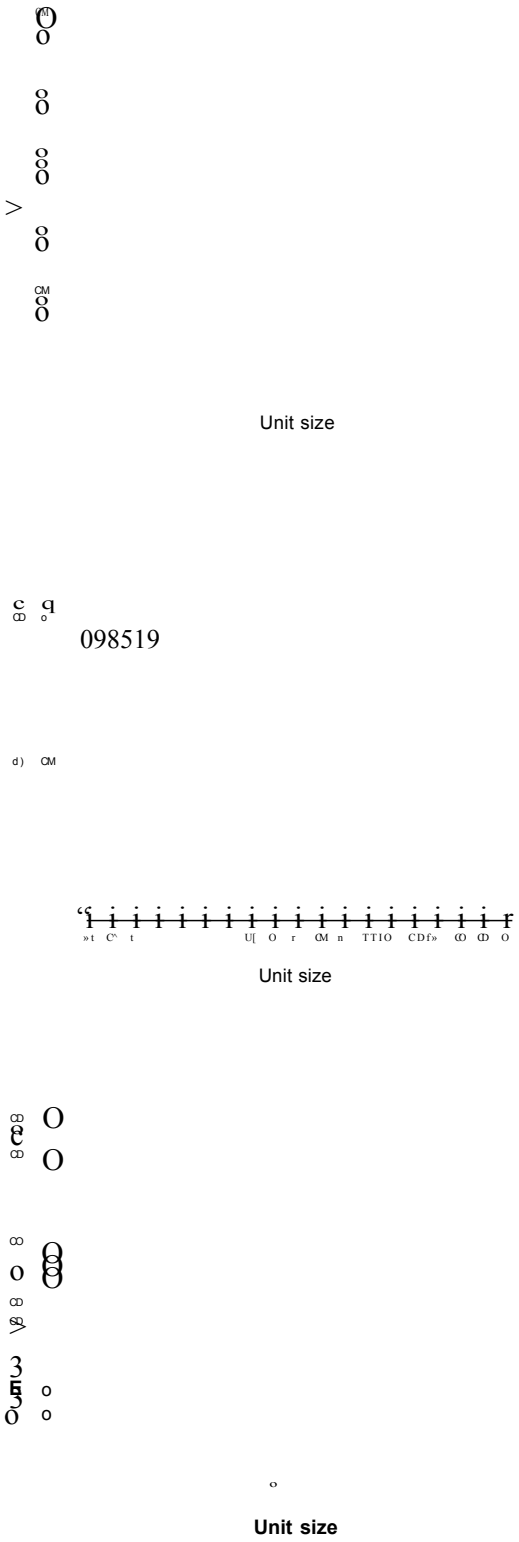


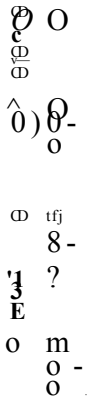
Figure 4-11: cumulative error rates based on the number of hidden units for three folds (top plot), five folds (middle plot) and ten folds (bottom plot) in terms of uncontrolled factors.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Height	1	0.207	0.207	0.207
Work connected with radiation	2	0.113	0.113	0.113
Spontaneous abortions	3	0.096	0.096	0.097
Age at menarche	4	0.092	0.094	095
Age at menopause	6	0.092	0.092	0.092
Age at firs Pregnancy	7	0.091	0.092	0.092

Table 4-12: Error rate of cross validation according to the neural network trained on uncontrolled factors

4.3.5 Demographic and controlled factors

Fig.4-12 shows that two units are chosen to build the neural networks classifier. The weight variable leads to the best performance among these factors (Table 4-13), and the best reduction in error rate is obtained by the classifier trained on weight and breastfeeding as shown in Table 4-13. Weight, breastfeeding, length of breastfeeding, sporting activity, kinds of meat and socioeconomic factors are the variables used for further analysis. The error rate is 76.87 per cent for the three cross-validation procedures, so the unremarkable employee, age and educational level are excluded.



Unit size

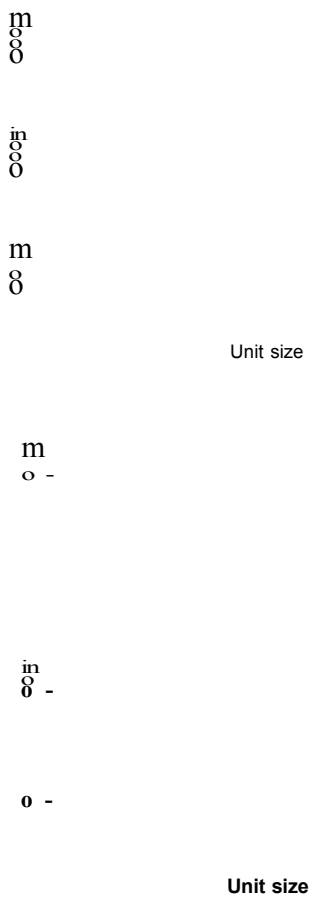


Figure 4-12: cumulative error rates based on the number of hidden units for three folds (top plot), live folds (middle plot) and ten folds (bottom plot) in terms of demographic and control factors.

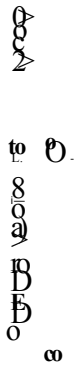
Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Breastfeeding	2	0.169	0.169	0.169
Length of Breastfeeding	3	0.138	0.137	0.137
Sport	4	0.111	0.110	0.110
Kind of Meat	5	0.086	0.085	0.085
Socio Economic	6	0.077	0.078	0.078
Employee	7	0.077	0.076	0.077
Age	8	0.073	0.075	0.073
Education level	9	0.072	0.074	0.073

Table 4-13 :Error rate of cross validation according to the logistic models trained based on demographic and controlled factors.

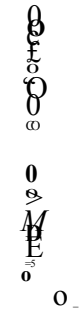
4.3.6 Uncontrolled and health factors

Fig. 4-13 makes it clear that two units are best for constructing a selection classifier. The family history variable is the most accurate of the uncontrolled and health factors,

as shown in Table 4-14. With an accuracy of 91.6 per cent for the all procedures of cross-validation, family history and height provide the best combination of variables, a performance that improves further to 93.1 per cent after work connected with radiation is added. The error rate does not drop further when the other variables are added to the models 4, 5 and 6 see Table 4-14.



Unit size



Unit size

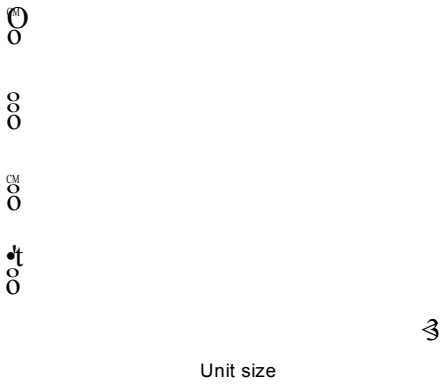


Figure 4-13: cumulative error rates based on the number of hidden units for three folds (top plot), five folds (middle plot) and ten folds (bottom plot) in terms of health and uncontrolled factors.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Height	2	0.084	0.084	0.084
Work connected with radiation	3	0.069	0.069	0.069
Inherited diseases	4	0.068	0.067	0.067
Spontaneous abortions	5	0.056	0.057	0.060
Other diseases	6	0.056	0.054	0.054

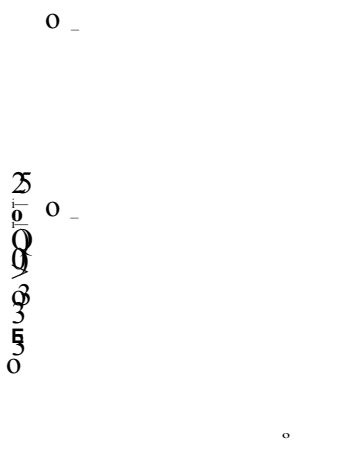
Table 4-14: Error rate of cross validation according to the neural network models trained on best uncontrolled and health factors

4.3.7 All factors

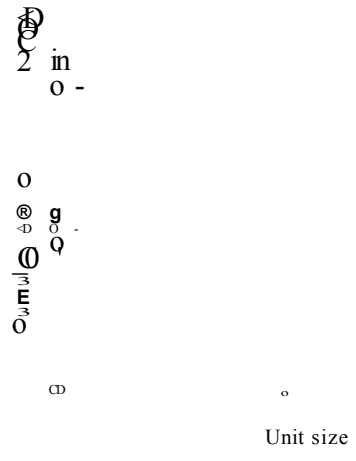
Two hidden units is the optimal number for all factors (Fig 4-14). With respect to the last stage of analysis, the best neural network classifier is shown in Table 4-15. Family history is the simple classifier that best discriminates between patient and control. The error rate drops significantly when height is added to the model, as demonstrated by Table 4-15. The final neural network classifier based on the variable selection algorithm shows a very low error rate using the five variables of family history, height, length of breastfeeding, work connected with radiation and breastfeeding respectively. In short, according to the three procedures of cross-validation, the error rate is about 3.7 per cent for classifier based on the five variables, while it is about 17.76 per cent using the simple model.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Height	2	0.084	0.084	0.084
Length of Breastfeeding	3	0.069	0.069	0.069
Work connected with radiation	4	0.046	0.048	0.047
Breastfeeding	5	0.037	0.040	0.039
Sport	6	0.034	0.034	0.034
Weight	7	0.034	0.030	0.031
Kind of Meat	8	0.028	0.029	0.0261

Table 4-15 Error rate of cross validation according to the neural network models trained on the all important factors



Unit size



Unit size

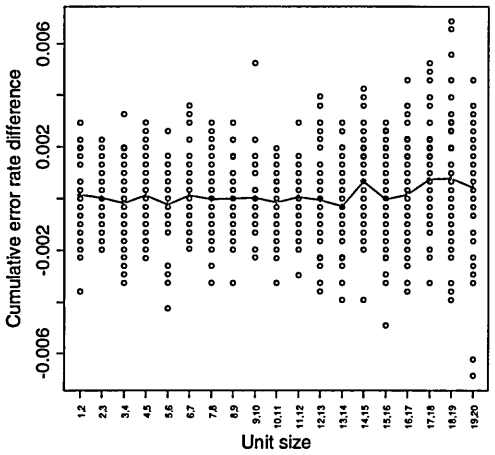


Figure 4-14: cumulative error rates based on the number of hidden units for three folds (top plot), five folds (middle plot) and ten folds (bottom plot) in terms of all factors.

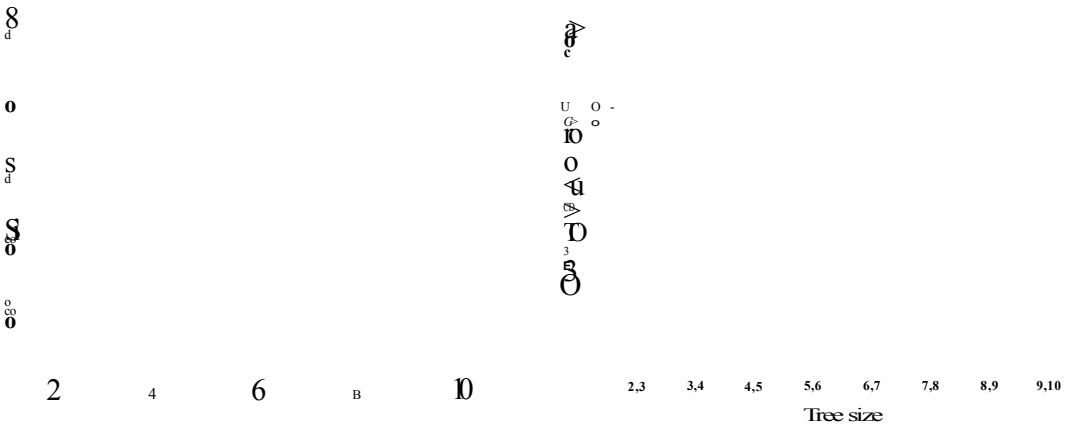
4.4 Decision Tree Method

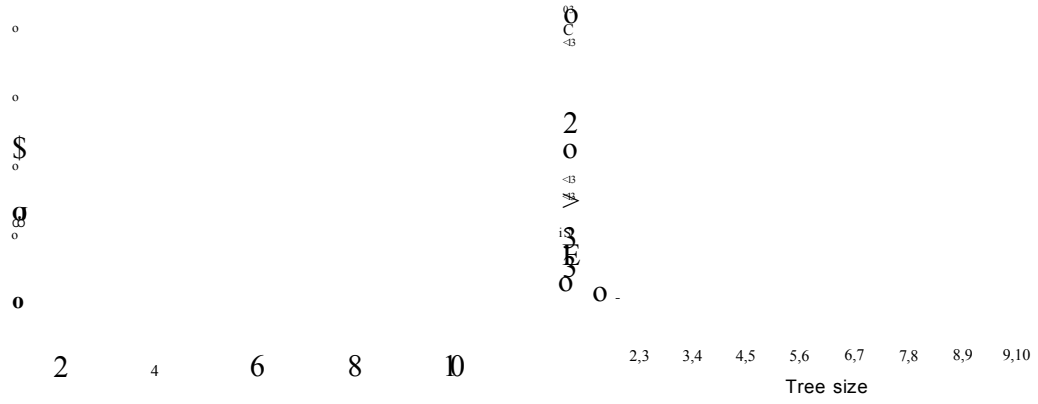
For the decision tree, the best size with the lowest error rate will be determined by constructing a large number of trees so that the complexity of the resulting tree classifier will less and the validation error low. To achieve this goal, the analysis was repeated 50 times using the cross-validation approach where K= {3, 5, 10}. The average error rate for training and validation is computed for each repetition. Trees were grown each time, the differences in error rates determined for two successive trees with different sizes and the cumulative sum for the error rate difference computed. Using this strategy it can be seen that, if the error rate for two successive trees decreases considerably, the cumulative curve rises sharply; also, if the curve rises slowly, the reduction in error rate will do likewise, resulting in a large tree. When the error rate becomes large the curves dips, and hence the tree should not be grown.

4.4.1 Demographic factors

Based on 50 three-folds repetitions Fig. 4-15 shows that the training and test error rates for demographic factors drop sharply when the tree size becomes six; this can be seen for three and five-folds. The plot for cumulative error rate difference makes it clear that the curve drops when the tree has seven terminal nodes. There is a slight difference for ten-folds, however: size seven leads to a little improvement. However, this does not result in a low error rate, as given by the corresponding cumulative rate. The validation error rate begins rising when tree size is larger than seven. When the size of tree becomes large, say more than six nodes, the validation error rate either does not decrease or it rises. In fact the seven-node performance using ten-folds does not result in

a marked improvement, so the performance of the accuracy at size six has been taken as optimal. This results in an allocation that is 69.9 per cent correct. The construction of the decision tree is shown in Figure 4-16. This plot makes it clear that the socioeconomic condition is the most important variable in the development of breast cancer, followed respectively by the level of education, age and employment.





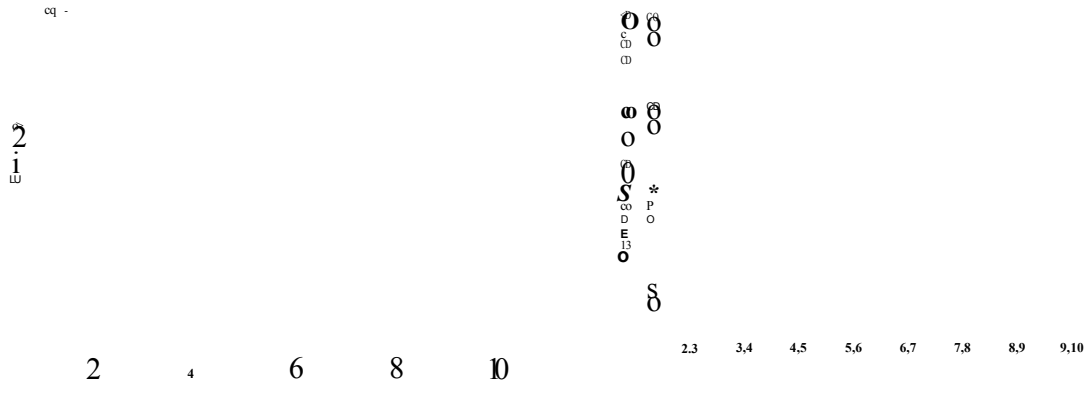


Figure 4-15: L.H.S shows error rate for different sizes of tree trained on demographic factors where green line is training error whereas red line is validation error. R.H.S shows the cumulative error rate difference. Notice that top panel shows three folds, middle panel shows five folds and bottom panel shows ten folds

Figure 4-16: Decision tree demographic factors using the best size for 3,5 and 10 folds.

4.4.2 Controlled factors

The results of tree pruning in terms of controlled factors are shown in Fig 4-17. According to these plots, the validation samples show that the training and test errors become closer as the tree size becomes smaller, whereas the situation is in fact the reverse. It is clear for all cross-validation folds that the error rates decrease as tree size increases. But the improvement in tree classifier is not as great as the complexity is high. The plots show that accuracy at size seven delivers the optimal performance, resulting in an allocation that is 88.06 per cent correct. Since this size provides terminal nodes for trees with the same label in the same branch, these nodes can be combined. Snipping these nodes thus produces six terminal nodes. The tree has accordingly been constructed as shown in Fig 4-18. It is clear that weight is the most important variable in the development of breast cancer, followed by sporting involvement and length of

breastfeeding.

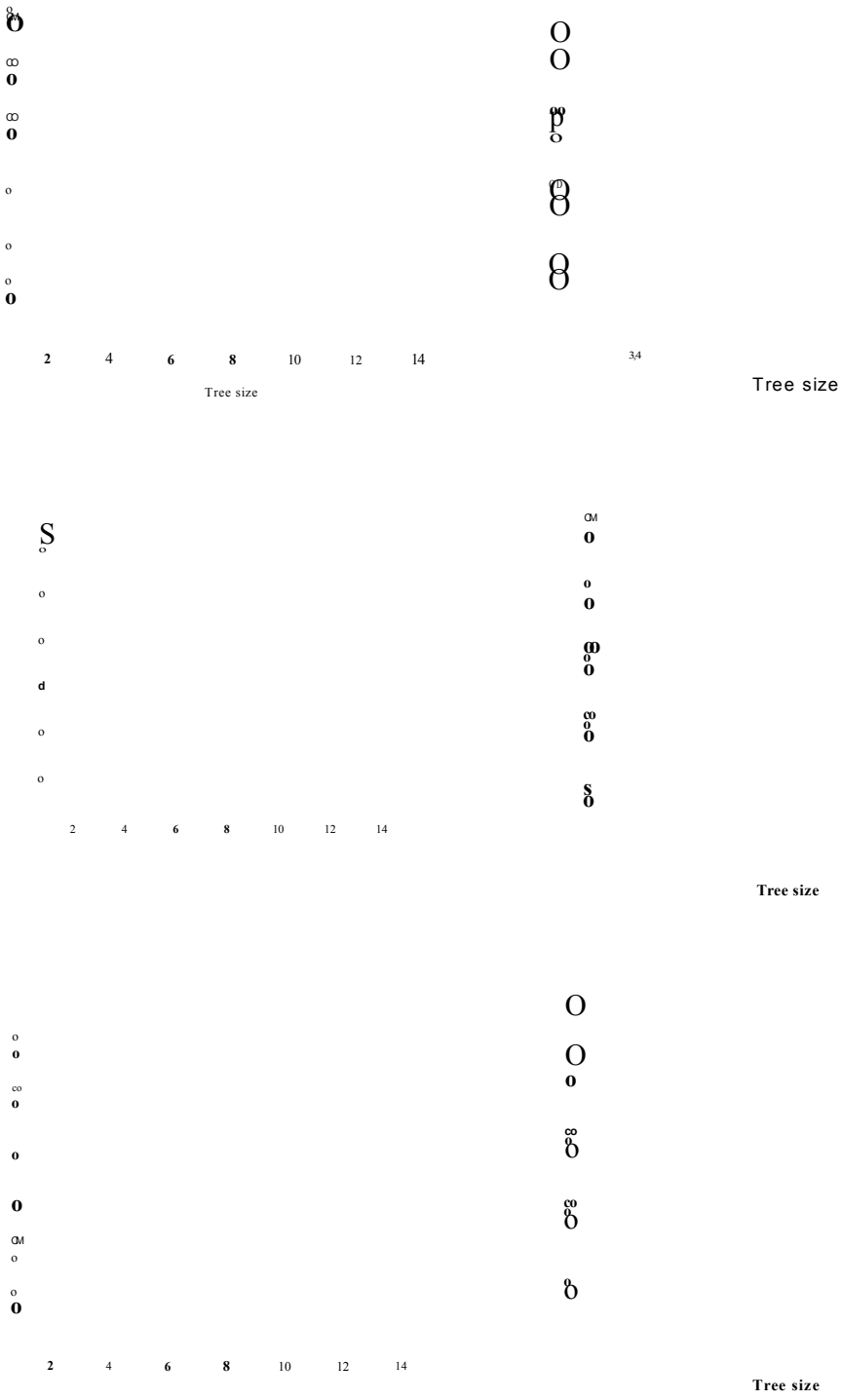


Figure 4-17:L.H.S shows error rate for different sizes of tree trained on controlled factors where green line is training error whereas red line is validation error. R.H.S shows the cumulative error rate difference. Notice that top panel shows three folds, middle panel shows five folds and bottom panel shows ten folds

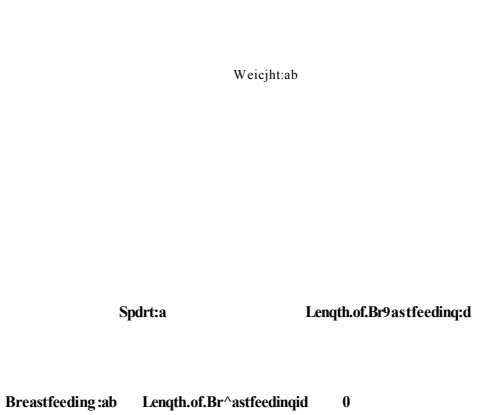


Figure 4-18: Decision tree for controlled factors using the best size for 3,5 and 10 folds.

4.4.3 Uncontrolled factors

The results for uncontrolled factors are given in Fig 4-19 the training and test error rates are the same for all folds except for five-folds, which differ somewhat at the tree of size three. When the tree becomes much bigger - say, more than six - the error will be high. This trend becomes much clear as the number of folds increases (see the plots for cumulative error rate difference). For all folds the accuracy at size six has been taken as the optimal performance, resulting in an allocation that is 90.67 per cent correct. The tree was constructed accordingly. Fig 4-20 this plot makes it clear that height is the most important variable in the development of breast cancer, followed by working with radiation and spontaneous abortion.

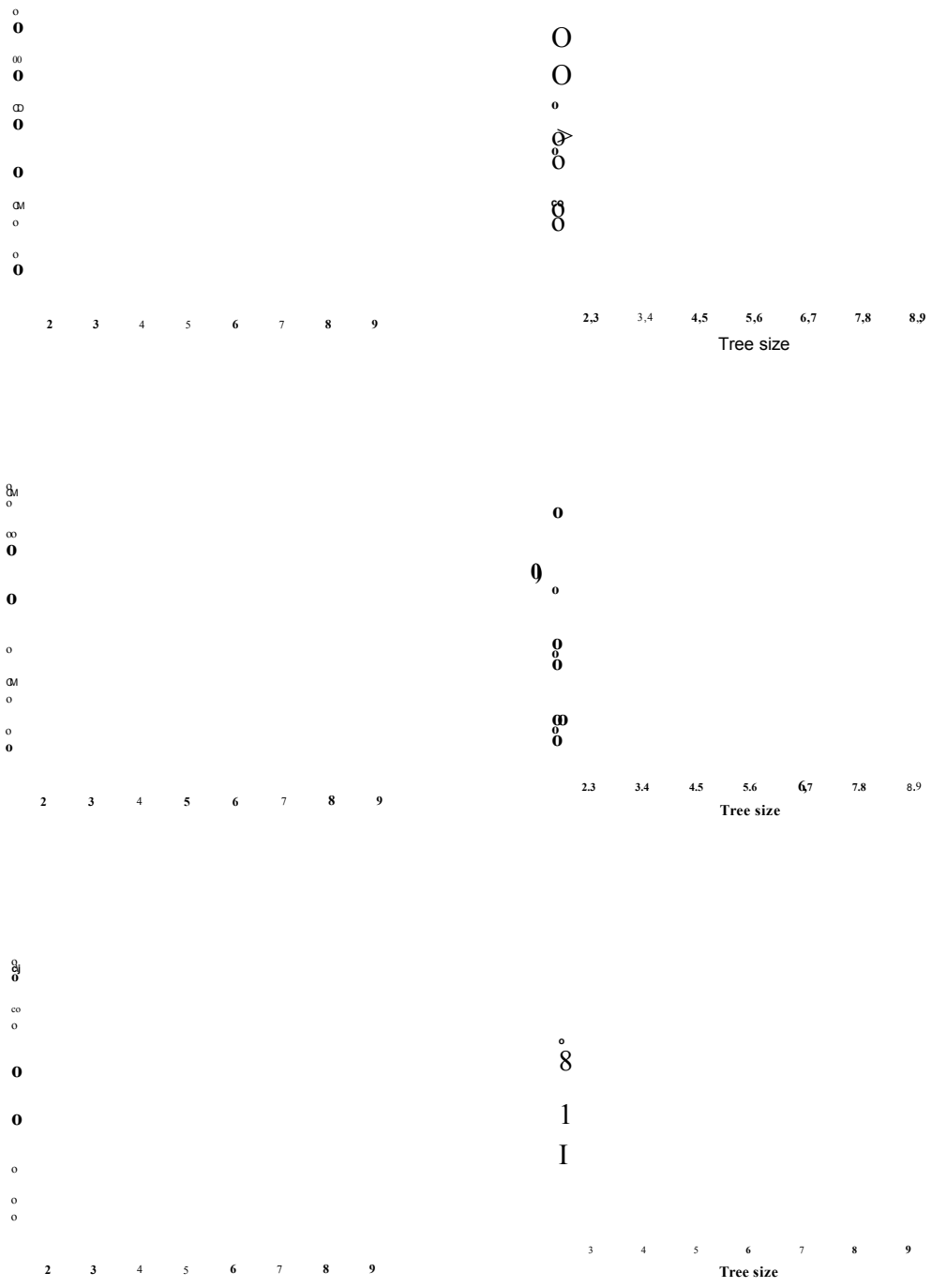


Figure 4-19 : L.H.S shows error rate for different sizes of tree trained on uncontrolled factors where green line is training error whereas red line is validation error. R.H.S shows the cumulative error rate difference. Notice that top panel shows three folds, middle panel shows five folds and bottom panel shows ten folds

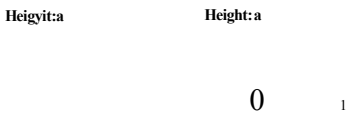
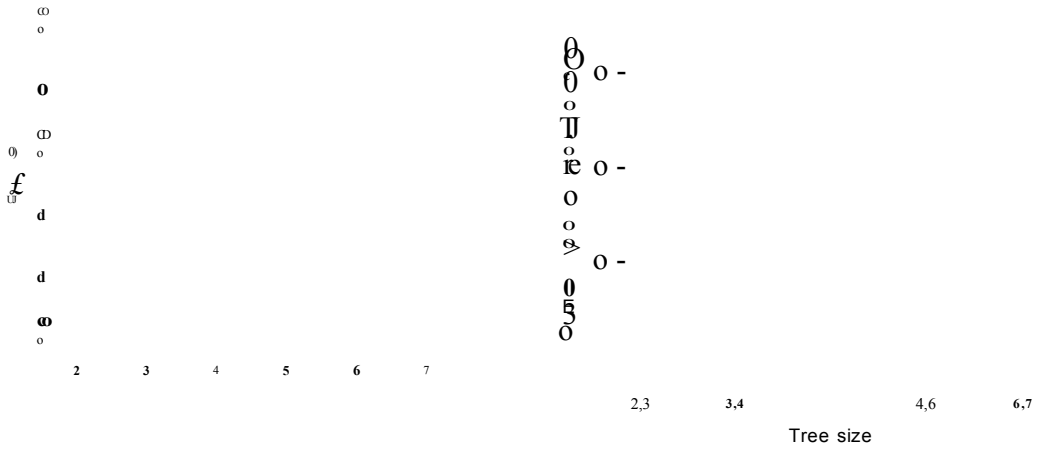


Figure 4-20: Decision tree for uncontrolled factors using the best size for 3,5 and 10 folds.

4.4.4 Health factors

For these factors, the training and test error rates for health factors drop sharply when the tree size is four, as shown in Fig 4-21. On the other hand, the difference in trend between training and test error is obvious in the plot for cumulative error rate difference. All plots confirm that tree sizes any larger than this do not affect performance. The size four decision tree displays the optimal performance of 87.75 per cent and was accordingly constructed as shown in Fig 4-22 .It is clear that family history is the most important variable in the development of breast cancer, followed by other diseases and inherited diseases.



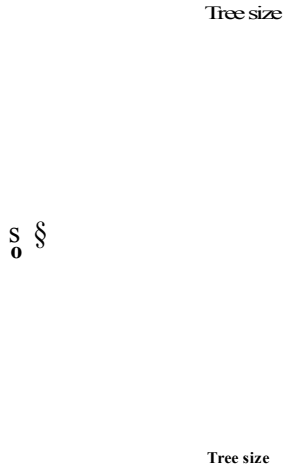


Figure 4-21: L.H.S shows error rate for different sizes of tree trained on health factors where green line is training error whereas red line is validation error. R.H.S shows the cumulative error rate difference. Notice that top panel shows three folds, middle panel shows five folds and bottom panel shows ten folds

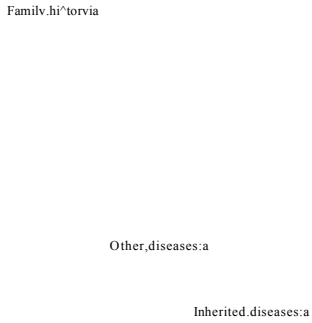
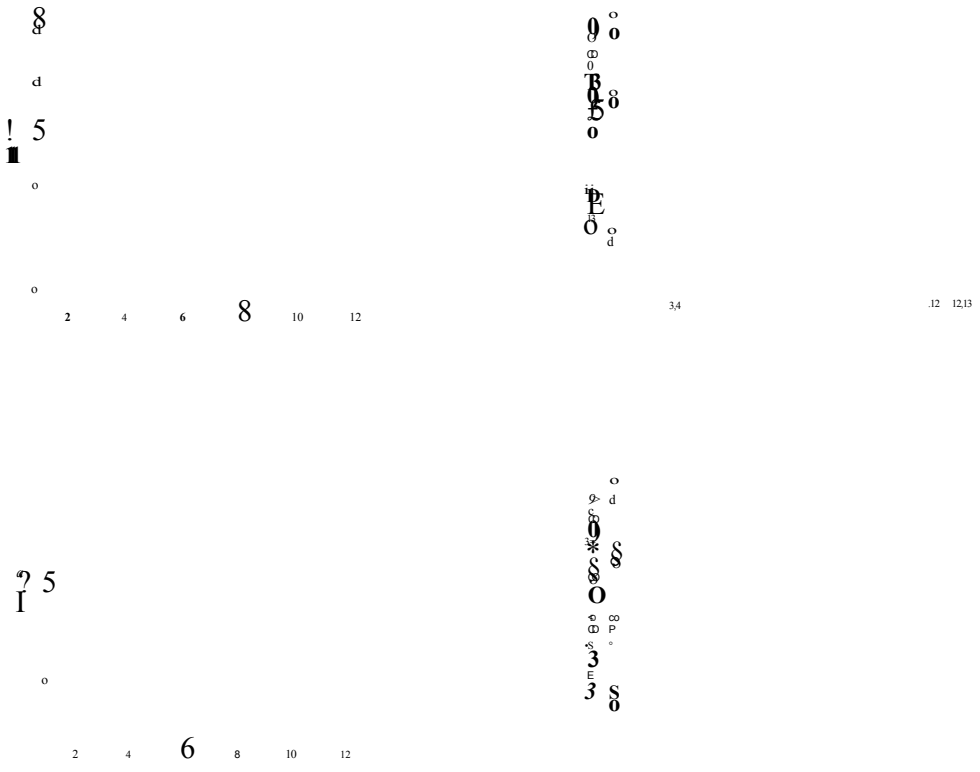


Figure 4-22: Decision tree for health factors using the best size for 3,5 and 10 folds.

4.4.5 Demographic and controlled factors

The error rate curves obtained by the all folds are somewhat similar see Fig 4-23. A tree of size 12 has the lowest value error, but the difference between this error rate and that for a tree of size five is small. Based on all the procedures, the splitting terminates at five terminal nodes, a size that results in an allocation that is 84.39 per cent correct. The tree was accordingly constructed as shown in Fig4-24 this plot shows that weight is the most important variable in the development of breast cancer, followed by length of breastfeeding and sporting activity.



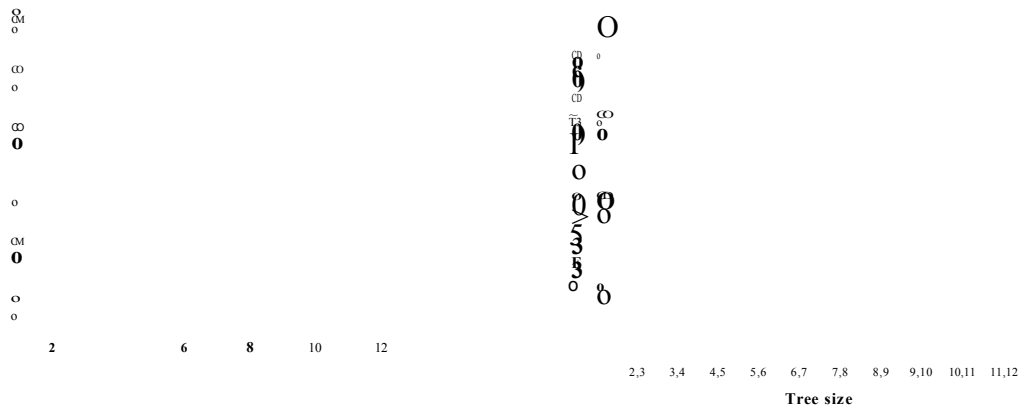


Figure 4-23:L.H.S shows error rate for different sizes of tree trained on demographic and control factors where green line is training error whereas red line is validation error. R.H.S shows the cumulative error rate difference. Notice that top panel shows three folds, middle panel shows five folds and bottom panel shows ten folds

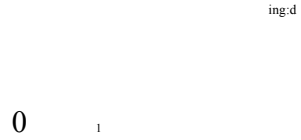


Figure 4-24: Decision tree for demographic and control factors using the best size for 3,5 and 10 fold.

4.4.6 Uncontrolled and health factors

Figure 4-25 shows that the training and test error rates for uncontrolled and health factors drop sharply when the tree size reaches four. In terms of test sets, the figure displays that any greater size than this does not result in a markedly lower error rate, and this has consequently been adopted as the size that results in the best performance. The resulting decision tree is shown in Fig 4-26 it is clear from this plot that height is the most important variable for developing breast cancer, followed by family history and height.

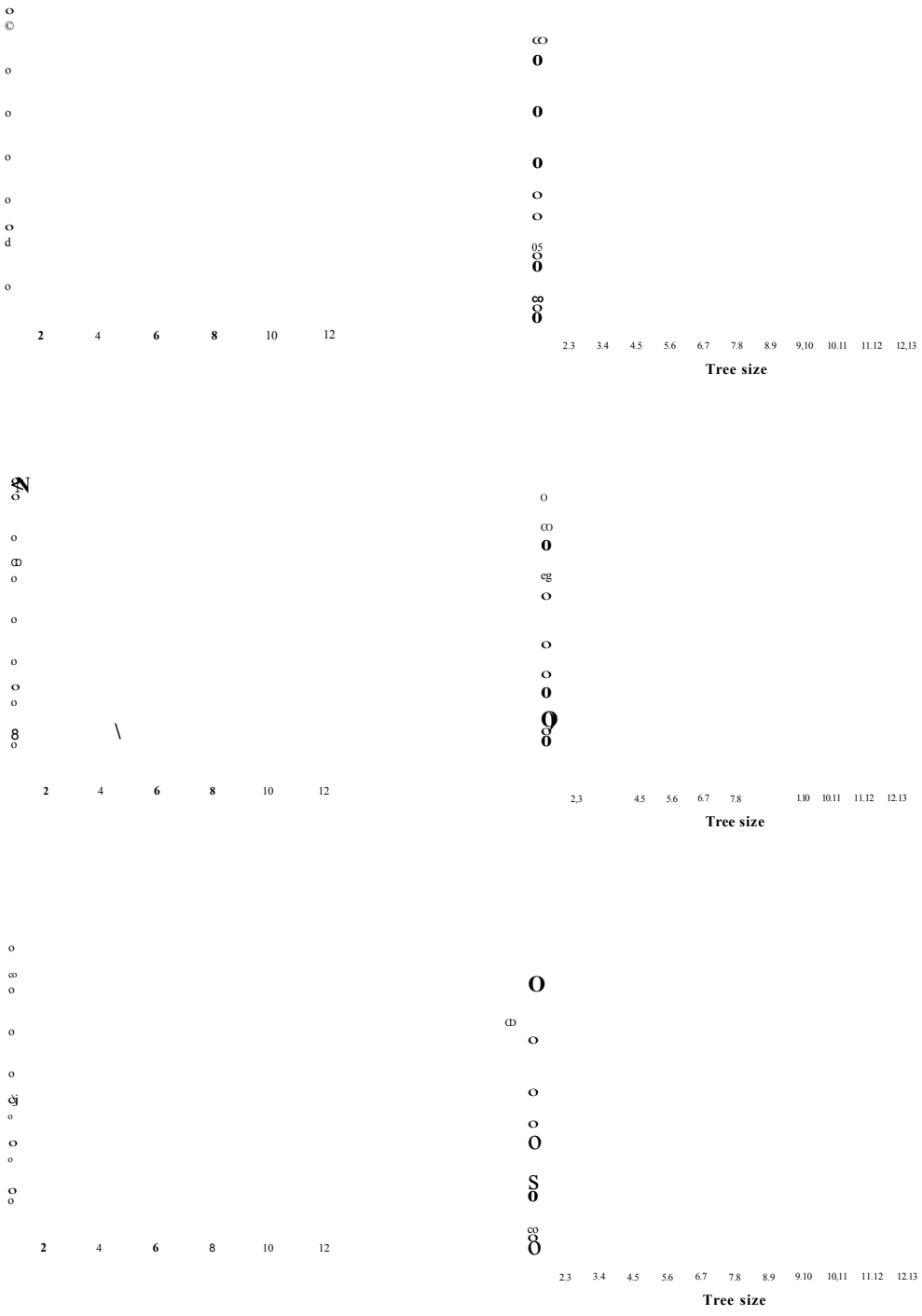
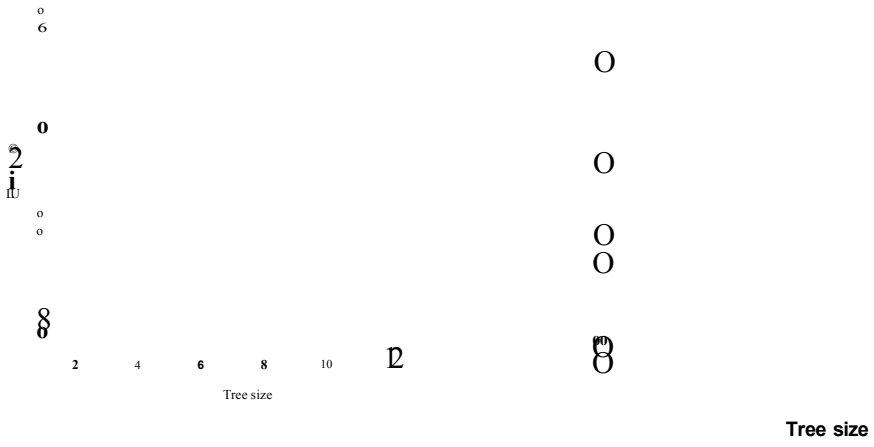


Figure 4-25:L.H.S shows error rate for different sizes of tree trained on health and uncontrolled factors where green line is training error whereas red line is validation error, R.H.S shows the cumulative error rate difference. Notice that top panel shows three folds, middle panel shows five folds and bottom panel shows ten folds

Figure 4-26: Decision tree for health and uncontrolled factors using the best size for 3,5 and 10 folds

4.4.7 AH factors

For the final model, Fig 4-27 shows that the size of decision tree resulting in the best performance is seven. The results show that tall women are more prone to the disease see Fig 4-28 .The tree diagram given in the same figure demonstrates that height, family history, weight, length of breastfeeding and sporting activity are respectively the most important variables for classifying the study individuals.



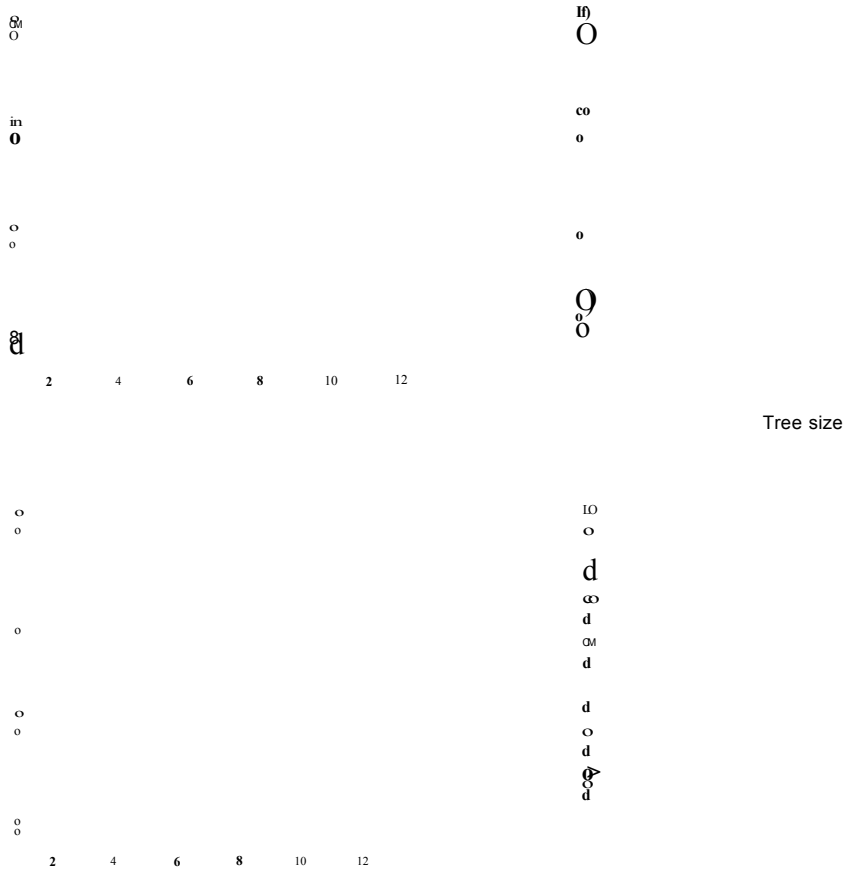


Figure 4-27: L.H.S shows error rate for different sizes of tree trained on all factors where green line is training error whereas red line is validation error. R.H.S shows the cumulative error rate difference. Notice that top panel shows three folds, middle panel shows five folds and bottom panel shows ten folds.

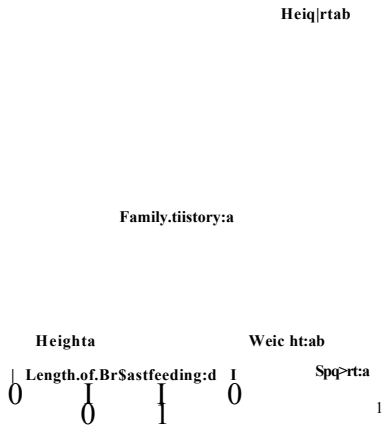


Figure 4-28: Decision tree for all factors using the best size for 3, 5 and 10 folds

By using complexity parameters (CPs) in order to avoid over-fitting resulting from growing decision trees, Table 4-16 shows the error rate using K-folds cross-validation for each combination of factors. For each combination, the CP values of 0.5 and 0.2 lead to very similar error rates across all the investigated folds. On the other hand, the error rate falls considerably when the CP is set to be .001. This reduction in the error rate does not differ much across the folds.

Factors Combination	**CP**	**Error rate**					
		3 folds		5 folds		10 folds	
		test	train	test	train	test	train
Demographic	0.50	0.490	0.488	0.488	0.488	0.488	0.488
	0.20	0.424	0.412	0.424	0.411	0.413	0.401
	0.001	**0.266**	**0.251**	**0.264**	**0.250**	**0.259**	**0.250**
Controlled	0.50	0.209	0.209	0.209	0.209	0.209	0.209
	0.20	0.209	0.209	0.209	0.209	0.209	0.209
	0.001	**0.087**	**0.071**	**0.085**	**0.071**	**0.084**	**0.072**
Uncontrolled	0.50	0.209	0.207	0.208	0.207	0.207	0.207
	0.20	0.209	0.207	0.208	0.207	0.207	0.207
	0.001	**0.094**	**0.083**	**0.092**	**0.084**	**0.091**	**0.084**
Health	0.50	0.177	0.177	0.177	0.177	0.177	0.177
	0.20	0.177	0.177	0.177	0.177	0.177	0.177
	0.001	**0.132**	**0.132**	**0.132**	**0.132**	**0.132**	**0.132**
Demographic and controlled	0.50	0.209	0.209	0.209	0.209	0.209	0.209
	0.20	0.209	0.209	0.209	0.209	0.209	0.209
	0.001	**0.091**	**0.073**	**0.089**	**0.074**	**0.089**	**0.074**
Uncontrolled and health	0.50	0.181	0.178	0.180	0.178	0.178	0.177
	0.20	0.181	0.178	0.180	0.178	0.178	0.177
	0.001	**0.063**	**0.05**	**0.063**	**0.054**	**0.063**	**0.055**
All factors	0.50	0.181	0.178	0.180	0.178	0.178	0.177
	0.20	0.181	0.178	0.180	0.178	0.178	0.177
	0.001	**0.045**	**0.036**	**0.041**	**0.034**	**0.039**	**0.034**

Table 4-16: Error rates for the best models of decision tree using cross-validation for different values of CP

Using CPs can build trees with somewhat lower errors than is possible with just rule of thumb selection. However, decision trees based on CPs can have larger terminal nodes than the rule of thumb used in Table 4-16. CP-based performance does not differ significantly from that obtained by rule of thumb. Moreover, CP seems to be affected by the size of the training sample, where error rates decline as the numbers of folds increase, unlike rule of thumb, where the training size does not affect performance.

4.5 Summary

In this chapter detailed analysis is carried out for logistic regression, neural network and decision tree. Thus, the third objective of the study (Provide breast cancer findings from the same cultural background, Libya, as a comparative input into the global geographical distribution of breast cancer) is achieved.

We found through logistic regression and neural network that weight, sporting activity, length of breast feeding, kind of meat consumed, breast feeding, height, family history and work related to radiation are important factors which trigger or cause breast cancer. Accuracy for LR and NN was 96.3. In the case of the decision tree, the important factors we found contributing most to breast cancer were height, family history, weight, length of breast feeding, and sporting activities. Accuracy for DT was 93.4.

Chapter 5 : Further Analysis of Data: 2-Stage Modelling

5.1 Introduction

A difference between neural network and decision tree modelling is that the latter more easily determines which factors are plausible classifiers, but neural networks can use all factors fed into them. This suggests the desirability of a hybrid 2-stage model. In this hybrid, a decision tree is used to decide the best factors to use and these factors are then passed into neural network architecture in order to produce a classification result.

Knowing also that logistic regression results are easier to interpret structurally than are neural network results, a second hybrid uses the same factors identified in the decision tree stage in a logistic regression model. The results of the base models (the decision tree, single neural network and logistic regression models already considered in Chapter 4) are then compared to the hybrid results.

5.2 Result DT Passed in to a NN

5.2.1 Demographic factors

Regarding the results of three, five and ten folds, for the variable selection algorithm the classifier trained on the socioeconomic variable leads to the highest performance. For models consisting of two variables, one of them should be the socioeconomic, one the greatest reduction in error rate is obtained by the classifier trained on the socioeconomic and educational level variables as shown in Table 5-1 By following the same process the socioeconomic, educational level, employment and age variables were selected, giving the best performance with the lowest error rate (26.3 per cent).

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Socio Economic	1	0.386	0.386	0.386
Education, level	2	0.355	0.355	0.355
Employee	3	0.310	0.310	0.310
Age	4	0.263	0.268	0.263

Table 5-1: Error rate of cross validation according to the neural network models trained on demographic factors.

5.2.2 Controlled Factors

Regarding controlled factors, variable selection shows that the classifier trained on weight leads to the best performance among the all simple classification models of neural network. When new classifiers with models consisting of two variables, one of

which is weight, are constructed, the greatest reduction in error rate is obtained by the classifier trained on weight and breastfeeding, as shown in Table 5.2. After several runs of the selection algorithm, weight, breastfeeding, length of breastfeeding and sports activities were selected as the variables for the next step of analysis, since they show the lowest error rate (11.2 per cent).

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Breastfeeding	2	0.169	0.169	0.169
Length of Breastfeeding	3	0.138	0.138	0.139
Sport	4	0.112	0.114	0.114

Table 5-2: Error rate of cross validation according to the neural network trained on controlled factors.

5.2.3 Uncontrolled Factors

With respect to uncontrolled factors, the height variable is the simple classifier that provides the highest performance. The construction of another classifier consisting of two variables, one of which is height, shows that the greatest reduction in error rate comes from the classifier trained on height and working with radiation, as shown in Table 5-3. Following the same procedure, height, working with radiation and miscarriages are the factors chosen for the further step of analysis; they demonstrate the best performance with the lowest error rate (9.6 per cent).

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Height	1	0.207	0.207	0.207
Work connected with radiation	2	0.113	0.113	0.113
Spontaneous abortions	3	0.096	0.096	0.096

Table 5-3: Error rate of cross validation according to the neural network trained on uncontrolled factors.

5.2.4 Health Factors

In terms of health factors, the selection of classifier based on a single variable shows that family history gives the highest performance, with an error rate of 17.8 per cent as shown in Table 5-4. New models consisting of two variables, one of which is family history, results in the greatest reduction in error rate being obtained by the classifier trained on family history and other diseases, while for models consisting of three variables, family history, other diseases and inherited diseases gives the best

performance with the lowest error rate(13.2 per cent).

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Other diseases	2	0.144	0.144	0.144
Inherited diseases	3	0.132	0.132	0.132

Table 5-4: Error rate of cross validation according to the neural network model trained on health factors

5.2.5 Demographic and controlled factors

For the best demographic and controlled factors, variable selection shows that weight leads to the highest performance (Table 5-5) .For models consisting of two variables the greatest reduction in error rate is obtained by the classifier trained on weight and sporting activity. These two variables together with length of breast feeding were selected for the next step of analysis. The error rate is about 12.40 per cent.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Sport	2	0.180	0.180	0.180
Length of Breastfeeding	3	0.124	0.126	0.125

Table 5-5: Error rate of cross validation according to the logistic models trained based on demographic and controlled factors

5.2.6 Uncontrolled and health factors

Regarding selected uncontrolled and health factors, the family history variable gives the highest accuracy (Table 5-6). With an accuracy of 91.6 per cent for all cross-validation procedures, height gives the best performance when one variable is added to family history.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Height	2	0.084	0.084	0.084

Table 5-6: Error rate of cross validation according to the neural network models trained on best uncontrolled and health factors.

5.2.7 All factors

With respect to the last stage of analysis, the best neural network classifier is shown in Table 5-7. For a simple classifier, family history best helps to distinguish between

patients with and without the disease. The error rate drops significantly when height is added to the model. The final model classifier based on the variable selection algorithm results in very low error rates when using the five variables of family history, height, length of breastfeeding, weight and sporting activity respectively, with a minimum error rate of about4.6 per cent (the simple model equivalent is about 17.8 per cent).

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Height	2	0.084	0.084	0.084
Length of Breastfeeding	3	0.069	0.069	0.069
Weight	4	0.057	0.055	0.055
Sport	5	0.048	0.046	0.049

Table 5-7: Error rate of cross validation according to the neural network models trained on the all

5.3 Result of DT passed into a LR

5.3.1 Demographic factors

The selection results for the demographic factor shown in Table 5-8 demonstrate that the socioeconomic variable most helps to distinguish between the two groups for a simple model consisting of one variable. By adding the other important variables using the selection algorithm, the selection process is terminated at the socioeconomic, educational level, age and employment variables, since it is believed that no important reduction in the error rate can be further achieved.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Socio Economic	1	0.386	0.386	0.386
Education, level	2	0.355	0.355	0.355
Age	3	0.346	0.348	0.348
Employee	4	0.313	0.315	0.317

Table 5-8: Error rate of cross validation according to the logistic models trained on models demographic factors.

5.3.2 Controlled Factors

Based on controlled factors, the results given in Table 5-9 show that weight gives the highest performance for simple models (about 79 per cent); this accuracy is considerable for a model based on one independent variable. For the best logistic regression models based on two and three variables, the results show that breastfeeding and length of breast feeding are the most important respectively, and that for the model

based on three variables the error rate has dropped to 13.7 per cent. By adding the other important variables using the selection algorithm, the selection process is terminated at the four variables of weight, breastfeeding, length of breastfeeding and sport in activity.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Breastfeeding	2	0.169	0.169	0.169
Length of Breastfeeding	3	0.137	0.137	0.137
Sport	4	0.113	0.112	0.111

Table 5-9: Error rate of cross validation according to the logistic models trained on controlled factors

5.3.3 Uncontrolled Factors

Regarding uncontrolled factors as shown in Table 5-10 height is the most important variable when distinguishing between the groups, followed by work connected with radiation and miscarriages. The error rate drops from 20.7 per cent for the simple model to 9.2 per cent for the one based on those three variables.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Height	1	0.207	0.207	0.207
Work connected with radiation	2	0.113	0.113	0.113
Spontaneous abortions	3	0.092	0.092	0.092

Table 5-10: Error rate of cross validation according to the logistic models trained on uncontrolled factors.

5.3.4 Health Factors

The results in Table 5-11 show that, for a simple model, family history of breast cancer is the most important variable; it is consequently the main one responsible for causing the disease, with a prediction accuracy of 81.3 per cent. Other diseases and inherited diseases variables do not contribute to the classification performance.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Other diseases	2	0.178	0.178	0.178
Inherited diseases	3	0.178	0.178	0.178

Table 5-11: Error rate of cross validation according to the logistic models trained on health factors

5.3.5 Demographic and controlled factors

Based on the selection results of demographic and controlled factors shown in Table 5-12, weight is the first variable entered into the model; its error is 20.9 per cent using three, five and ten folds. This rate drops significantly to 18.1 per cent when sporting activity is entered, and continues gradually to decline on addition of the other important variables found with the selection algorithm. The selection process is terminated at the three variables of weight, sporting activity and length of breast feeding. The best controlled factors will be used to select the final model. Based on this result, the effect of demographic factors, probably, pass through the controlled factors.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Weight	1	0.209	0.209	0.209
Sport	2	0.181	0.181	0.181
Length of Breastfeeding	3	0.123	0.123	0.123

Table 5-12: Error rate of cross validation according to the logistic models trained on best demographic and control factors

5.3.6 Uncontrolled and health factors

The results of the variables selection algorithm shown in Table 5-13 demonstrate that family history comes first, with an error rate of 17.8 per cent according to the three types of folds. This factor is consequently considered to be a major factor causing breast cancer. In the second model by height, the error rate falls to 8.4 per cent, which is a good reduction. Hence the selection process is terminated at these two variables. The reduction in error rate to which they give rise is relatively low according to the three types of cross-validation.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Height	2	0.084	0.084	0.084

Table 5-13: Error rate of cross validation according to the logistic models trained on best uncontrolled and health factors.

5.3.7 All factors

In order to increase the classification performance of the best variables obtained from Table 5-12 and Table 5-13 that are used in the final model, Table 5-14 shows the greatest contribution to the classification. Family history and height each have an error

rate of 8.4 per cent. The length of breastfeeding, from uncontrolled factors, when added to the variables already selected for this model, reduces the error rate to 6.9 per cent. The best performance of 95.6 per cent is achieved by adding weight and sporting activity to these variables.

Variable	Number of variables in the model	Error rate		
		Three folds	Five folds	Ten folds
Family history	1	0.178	0.178	0.178
Height	2	0.084	0.084	0.084
Length of Breastfeeding	3	0.069	0.069	0.069
Weight	4	0.053	0.053	0.052
Sport	5	0.046	0.044	0.045

Table 5-14: Error rate of cross validation according to the logistic models based trained on the all important factors.

5.4 Summary

When we passed the most important factors from DT output and fed them into the neural network architecture there was an improvement in accuracy. It means NN is more accurate in interpreting the results of the DT, leading the researcher to conclude NN is more suitable to allocate the patient to the right group. Passing the output from DT to LR gives a better performance than NN; the accuracy reaches 95.6, which is excellent to predict the factors that are more important in contributing to the disease. The following chapter will be devoted to discussion of the analysis result.

Chapter 6 : Discussion of the Analysis

6.1 Introduction

In order to predict the risk of breast cancer in Libya, the previous Chapter 4 and 5 have analyzed the data collected according to our proposed strategy using three data mining techniques: logistic regression; neural network; and decision tree. The purpose of the current chapter is to discuss the result findings, show how the research questions have been addressed.

6.2 Logistic regression model

The cross-validation forward selection based on the three strategies arranges the factors of interest in the order of demographic, controlled, uncontrolled and health factors. Table 6-1 makes it very obvious that the three types of cross-validation result in the same classification errors. For the first stage, control variables exhibit the lowest error rate, followed by uncontrolled variables, with health and demographic factors coming third and fourth respectively. Demographic factors produce the lowest performance. The error rate drops dramatically as important variables selected from the factors are combined. The final model displays a very high accuracy of 96.3% - an excellent performance. This success is very important when understanding and controlling this virulent disease. The fact that three, five and ten-folds show similar error rates can be attributed to the large size of the folds, which is a consequence of the large size of the entire data set.

Factors	Three-folds	Five-folds	Ten-folds
Demographic	**0. 313**	**0.320**	**0.317**
Control	**0.085**	**0.085**	**0.085**
Uncontrolled	**0.092**	**0.092**	**0.092**
Health	**0.178**	**0.178**	**0.178**
Demographic & control	**0.085**	**0.085**	**0.085**
Uncontrolled & Health	**0.069**	**0.069**	**0.069**
All factors	**0.037**	**0.037**	**0.037**

Table 6-1: the error rate computed for best models of logistic regression using cross-validation technique

Based on the selection algorithm of the final classifier, the variables of weight, sporting activity, length of breastfeeding, kinds of meat consumed, breastfeeding, height, family history, and work connected with radiation are especially responsible for developing the disease. The chance of accruing breast cancer is seen to be much higher for older women than younger. It is much more likely to spread with age. Height in particular

increases the risk of contraction, especially for very tall women. Those who do not breastfeed are more susceptible to the disease, but any amount of breastfeeding reduces this risk. Women whose diet is based on fish and poultry are also safer. Women are strongly advised to work in places where there is no radiation. Sporting activity plays a significant role in lessening this risk.

6.3 Neural network

Table 6-2 shows, for the all folds of cross-validation samples, that the best neural network classifier performance is achieved firstly by controlled and then by uncontrolled factors, with demographic and health factors also producing reasonable levels of accuracy. After excluding unnecessary variables, the test error drops to 0.077 for demographic and control factors, a reduction that is not greatly different from that of controlled factors (0.085). By contrast, classification is improved when the set of important factors obtained by uncontrolled and health factors is involved in their construction.

Overall, the error rate is reduced to its lowest level of 3.7% to 3.9% for three to ten-folds when the final classifier is trained on the subset of the best factors. In fact, the differences in error rates derived from the cross-validation samples are very small, and are hence ignored. The numbers of weights will be the same for all folds. For each combination of factors, the number of weights will vary by the number of variables in each unit; the number of weights generally seems to be large if the number of variables is great.

Factors	Three folds		Five folds		Ten folds	
	Number of weights	Error rate	number of weights	Error rate	number of weights	Error rate
Demographic	35	0.264	35	0.269	35	0.265
Controlled	53	0.085	53	0.086	53	0.087
Uncontrolled	37	0.096	37	0.096	37	0.097
Health	11	0.132	11	0.132	11	0.132
Demographic and control	51	0.077	51	0.078	51	0.078
Uncontrolled and health	21	0.069	21	0.069	21	0.069
All factors	35	0.037	35	0.040	35	0.039

Table 6-2: the error rate computed for best models using cross-validation technique

Feed-forward neural networks are the same as logistic regression where no parametric assumptions are imposed on a dataset. Since the present data is in categorical form, neural networks are a very appropriate means of investigating the issue. The networks are based on input units connected to a layer of hidden units, which use a logistic function (activation function) in order to sum up the input units. The output units use the same form of activation function to allot individuals into their appropriate classes.

A back propagation learning algorithm is applied to estimate weights. The lowest number of hidden units resulting in high performance is two for the three, five and ten-folds experiments. This result is also returned for all combinations of factors. As a result, two units will be recommended where one hidden layer is used, irrespective of sample size and number of variables used for constructing neural network classifiers. As for logistic classifiers and neural networks methods based on cross-validation forward selection result in high performance, the methods have the same arrangement of the variables of interest. The selection results confirm that three, five and ten-folds give very similar accuracy.

6.4 Decision Tree

Decision trees are often built top-down – namely, decision trees are grown for a given dataset so that terminal nodes of decision tree are created. A test set is distributed to those

nodes so as to assess the classification capability of the decision tree. A large tree usually results in over-fitting with consequent redundancy of variables. This was dealt with in the present research by pruning decision trees using cross-validation trees. It was proposed that the cumulative error rate difference be found in order to determine the size of tree that demonstrates the highest performance. Plot seemed a better means of selecting the best size than plotting error rate against tree size. The shape of the cumulative curve is the same for the same set of variables if the three methods of cross-validation are used in some cases.

The results of the cross-validation experiments are presented in Table 6-3. The columns headed "-fold" show the average error rate and tree size for each factor. As can be seen, the uncontrolled variable gives the lowest error rate for the three-, five- and ten-folds: all of them give the same value of error rate (0.093) and size (six nodes). The highest error rate is returned by the demographic variable.

Factors	**Three folds**		**Five folds**		**Ten folds**	
	Error rate	Tree size	Error rate	Tree size	Error rate	Tree size
Demographic	0.310	6	0.310	6	0.310	6
Controlled	0.119	6	0.119	6	0.123	6
Uncontrolled	0.093	6	0.093	6	0.093	6
Health	0.130	4	0.130	4	0.130	4
Demographic and control	0.115	5	0.115	5	0.115	5
Uncontrolled and health	0.090	4	0.090	4	0.091	4
All factors	0.066	7	0.068	7	0.068	7

Table 6-3: the error rate computed for best models of decision tree using cross-validation technique

The combination of demographic and controlled factors is not markedly more accurate than controlled factors on its own. However, some little improvement is gained by the combination of uncontrolled and health factors, particularly for the three- and five-fold categories. The final models consisting of the most important factors obtained by the proposed plan demonstrate that the error rate drops to a very satisfying 6.6%.

The results show that the appropriate size of pruned trees is the same for the three procedures of cross-validation. The performance of the three procedures is also very similar, resulting in the same relative order of variable selection performance. The order of factor accuracy is the same as that of logistic classifiers. The graphic representation of

decision trees based on the final selection makes it apparent that overweight women who do not practice any kind of sport have been allocated into the cancer class.

It was observed that women with a positive family history are allocated to the breast cancer group when height increases and length of breastfeeding decreases. Overall, the decision tree classifier seems to be quite capable of allocating people to their appropriate classes, with an accuracy of 93.4%.

6.5 Summary Result DT Passed in to a NN

For the all folds of cross-validation samples, the best performance of neural network classifier is achieved by the uncontrolled and controlled factors respectively, as shown in Table 6-4 Health and demographic factors result in reasonable accuracy. An improvement in classification can be seen when the set of important factors obtained by uncontrolled and health factors is used to construct a classifier. Overall, the error rate reaches the lowest value (about 4.6% to 4.9% for three to ten folds) when the final classifier is trained on the best subset of the factors. In fact, the differences in error rates obtained by the cross-validation samples are very small, and hence are ignored here.

Factors	Three-folds	Five-folds	Ten-folds
Demographic	0.263	0.268	0.263
Control	0.112	0.114	0.114
Uncontrolled	0.096	0.096	0.096
Health	0.132	0.132	0.132
Demographic & control	0.124	0.126	0.125
Uncontrolled & Health	0.084	0.084	0.084
All factors	0.048	0.046	0.049

Table 6-4: the error rate computed for best models using cross-validation technique

6.6 Summary Result of DT passed into a LR

Table 6-5 makes it is very obvious that the three types of cross validation lead to the same classification errors. For the first stage, uncontrolled variables have the lowest

error rate followed by controlled ones. Health and demographic factors come third and fourth respectively. The demographic factor results in the lowest performance of all the factors. The final model demonstrates a very high accuracy of 95.6%. For this model, the important variables are, in order, family history, height, length of breastfeeding, weight and sporting activity.

Factors	Three-folds	Five-folds	Ten-folds
Demographic	0. 313	0.315	0.317
Control	0.113	0.112	0.111
Uncontrolled	0.092	0.092	0.092
Health	0.178	0.178	0.178
Demographic & control	0.123	0.123	0.123
Uncontrolled & Health	0.084	0.084	0.084
All factors	0.046	0.044	0.045

Table 6-5: the error rate computed for best models of logistic regression using cross-validation technique

6.7 Summary

According to the base decision tree model containing all factors, the performance was about 93.4% compared with the hybrid two-stage model. An improvement in classification can be seen when the set of those factors is passed into a neural network model, upon which the model's performance rises to 95.4%. When the same factors resulting from the decision tree are passed into a logistic regression model, the performance of the hybrid two-stage model achieves an accuracy of 95.6%, which falls short of that of the neural network and logistic regression models (96.3%), but is still acceptable as the best performance in distinguishing between who is and who is not prone to the disease. The last chapter provides a summary and evaluates the work, outlining its contributions and limitations, and makes recommendations for further work.

Chapter 7 : Conclusion and Recommendations

7.1 Introduction

This chapter focus mainly on the conclusions, limitations of the study, contribution to knowledge and policy worlds, and recommendations for future research and policy domain.

7.2 Contribution of this research

The literature review reveals many gaps in the current state of knowledge and understanding of the factors causing breast cancer particularly in developing countries such the African continent. This study proposed a new classification strategy making use of logistic, neural network and decision tree classifiers in order to predict the risk of breast cancer in Libya. The predictive strategy constitutes a major contribution to knowledge in that the research provides the use of a new data modelling strategy to select potentially predictive variables into different configurations of the three data mining techniques.

In this strategy, the variables of interest are allocated to the demographic, controlled, uncontrolled, health, and geographical groups, with each classifier being applied to each group. Due to the multidimensional dataset, high model complexity and poor performance are to be expected. One solution is to reduce the dimensionality. The present model enhances the performance of logistic and neural network classifiers by adding most of the important inputs (independent variables) using forward selection algorithms instead of selection algorithms; for a decision tree, the splitting rule retains only inputs providing small training error rates using deviance. This is not guaranteed to produce small error rates, so the pruning rule was applied in order to achieve this effect. This was how the performances of the three approaches were compared. Hence, the second research question was addressed *(How can we enhance breast cancer predicting models?).*

Selection of variables and pruning algorithms are evaluated by minimum error rate using the cross-validation algorithm. The algorithms are conducted for three, five and ten-folds so that a better evolution of performance can be achieved. Moreover, cross-validation of forward selection algorithms enables comparison and empirical selection of the classification rule. In other words, the dataset under consideration could itself

indicate which classification approach would enable the greatest predictive capability.

Although the three methods of classification concern the same variables and use cross-validation forward selection, analysis reveals that logistic and neural network classifiers are very little better than decision trees. For this data set, the number of folds all led to low error rates, so there was no particular preference. To save time, the writer recommends using two hidden units for neural networks based on one hidden layer. For decision trees, the difference in cumulative error rates was a good visual representation by which to choose the best size of tree.

In addition, according to the selection strategy explained in Chapter 3, the author arranged the predictors' factors regarding their importance for classification, as shown in Figure 6-1. It is very interesting to notice the relative importance of each predictor in terms of showing low error rates for each stage. It also provides the ability to observe how the contribution of some variables at any given stage can be absorbed by other variables at that same stage. Starting with six demographic and ten uncontrolled variables, a new model trained on just nine variables from both factors was established. For the new classifier trained on important variables of demographic and controlled factors, the five variables chosen from the controlled factors were entered into the model first, followed by demographic factors. It was thereby deduced that controlled factors can play a greater part in allocating objects into their appropriate classes than can demographic factors.

It should be remembered that the demographic factor is still an important predictor. Regarding health and uncontrolled factors, the selection procedure firstly chose family history from three variables belonging to the health factor, and then three variables selected from six belonging to the uncontrolled factor.

Generally, the final model based on the best variables retained from all the factors, family history, height, length of breastfeeding, work related to radiation, breastfeeding, kinds of meat consumed, sporting activity and weight are respectively the major variables that can distinguish between the two classes.

Demographic Factor	Controlled Factor	Uncontrolled Factor	Health Factor
1. Social-Economic 2. Education Level 3. Age 4. Employee Status 5. Marital Status 6. Gender	1. Weight 2. Breast Feeding 3. Length of Breast feeding 4. Sport 5. Kind of Meat 6. Age at last Pregnancy 7. Duration of Oral Contraceptives 8. Kind of Vegetables 9. Number of Children 10. Oral Contraception use	1. Height 2. Work Connection with radiation 3. Spontaneous Abortion 4. Age at Menarche 5. Age at Menopause 6. Age at first Pregnancy	1. Family History 2. Other Disease 3. Inherited Disease

1. Family History
2. Height
3. Work Connected with radiation
4. Spontaneous Abortion

1. Weight
2. Breast Feeding
3. Length of Breast feeding
4. Sport
5. Kind of Meat
6. Socio-Economic
7. Age
8. Employed Status
9. Education Level

1. Family History
2. Height
3. Length of Breast feeding
4. Work related with radiation
5. Breast Feeding
6. Kind of Meat
7. Sports
8. Weight

Figure 7-1: The arrangement of predictors using the selection strategy.

In spite of the results of current study almost agreeing with previous studies in terms of variables causing breast cancer, the study shows a contrary view about geographical regions. The regional factor is found to be an unimportant factor in classifying breast cancer in the specific case of Libya. Thus, the first research question was addressed *(Are breast cancer predictor factors same across the world?).* When different numbers of folds of cross-validation are used for a particular set of inputs, the performance of a

particular decision tree at difference sizes will sometimes be a little different. In other words, for the same set of observations measuring a set of inputs, the error rate for this set will not be exactly the same when evaluated by three, five or ten folds at particular sizes of decision tree.

The use of CP in building a decision tree results in a somewhat lower error rate than the method based on rule of thumb. Despite the lower error rate, decision trees based on the use of CP can have larger terminal nodes than those using rule of thumb. The respective performances of the two do not differ markedly. Moreover, while CP seems to be affected by the size training sample, where the error rates diminish as the numbers of folds increase, this is not the case for rule of thumb, where the training sizes do not affect performance.

It has been shown how statistical and machine-learning models can help doctors in Libyan hospitals better understand cancer risk factors in order to make an accurate diagnosis. All classification algorithms demonstrate high levels of discriminative accuracy and very similar performance levels. This may be attributed to the clarity of data set involved. In addition, all the methods yield a superior low error rate for the three procedures of cross-validation. All the classification methods have the potential to be used as decision support tools once they are integrated into clinical practice. It seems to be difficult to draw a general conclusion with respect to the superiority of one classification method over another on the basis of findings from this research's dataset.

The superiority of any classification method can only depend on the training dataset used with respect to size and classes overlapping. Even though work will be required to decide whether this performance is mainly attributable to the features of the present set or attributes of the present model itself, it is encouraging that this study's classification methods achieved an excellent measure of accuracy. Due to the conformity of the present results with previous ones, it can be accepted that this is the level at which machine learning classifiers will help hospitals improve their performance in early breast cancer prediction. A particular classifier generally possesses its own advantages; the selection of a classifier must be predicated on those advantages and on the purpose of the study.

In summary, the research makes the following contributions:

1. The major contribution to knowledge provided by the research is the use of a new data modelling strategy to select potentially predictive variables into different configurations of the three data mining techniques. Comparison between our results and those obtainable from the conventional LR, NN and DT models shows that our strategy out-performs the conventional variable selection.

2. To save time for other researchers, the author recommends using two hidden units for neural networks based on one hidden layer. In other words the best size of hidden units (those that minimise error rates in terms of the differences between cumulative error rates in neural network classifiers) is found.

3. The best number of terminal nodes (those that minimise error rates in terms of decision trees) is determined using differences in cumulative error rates.

4. The arrangement of predictors' factors according to their importance for classification and discriminating between the two groups of patients (with and without the disease) are listed.

5. The respective performances of the three classification methods are compared in order to recommend the most potent method of cancer predication.

6. We describe a simple procedure by which the strategy can be extended to other domain-partitioning models and highlight ways in which breast cancer research, prevention and cure stands to benefit in the long run.

7.3 Study Limitation

Like any other research, this study has a number of limitations. Two factors were excluded from the study: smoking and alcohol consumption. In Libyan culture women are not very likely to be smokers, and as a totally Muslim country, the consumption of alcohol is forbidden there. Using different numbers of folds of cross-validation for a particular set of inputs, the performance of a decision tree at difference sizes would be somewhat different. In other words, for the same set of observations measured on a set of inputs, the error rate for this set will not be exactly the same when evaluated by three, five, or ten folds at particular sizes of decision tree. The study has some methodological limitations. The quantitative approach, which here takes the form of a questionnaire,

does indeed provide a wide scope for investigation, but perhaps less so for detailed explanation, whereas a qualitative focus would be narrower but more exhaustive. The purposes of this research are best served by a quantitative study, especially in the light of the absence of other work in this area, although every effort has been made to ensure the inclusion of all relevant information regarding the patients' cases and treatment.

7.4 Recommendations for Further Research

The approach followed in this study can be extended in various ways to further research. In the case of ANNs, one should experiment with different activation functions. Although the logistic step function is used in this study, other similar functions mentioned in Chapter 3 could be used. Bishop (1995) for example applies the hyperbolic tangent activation function. It could be argued that similar classification results are achieved by ANNs and LR as a result of employing the logistic step function in this study.

For decision trees, it is well worth testing the decision tree approach on other important diseases such as haemophilia. It would also be interesting to test the decision tree approach on types of cancer other than breast cancer.

The study is limited to breast cancer risk factors in a small African country such as Libya, with a population of about 6,342,000. There is thus a need for further research in other developing countries with much larger populations in order to widen the range of investigation of the risk factors for this disease.

7.5 Conclusion

Our research shows how we can enhances the performance of logistic and neural network classifiers that is by adding most of the important inputs (independent variables) using forward selection algorithms instead of selection algorithms, for a decision tree, the splitting rule retains only inputs providing small training error rates using deviance. This is not guaranteed to produce small error rates; therefore the pruning rule was applied in order to achieve this effect.

The major variables that can distinguish between the two classes of patient, in final model based on the best variables retained from all the factors was ;family history, height, length of breastfeeding, work related to radiation, breastfeeding, kinds of meat consumed, sporting activity and weight are respectively.

In addition the author show variable selection is novel that is by arranged the predictors' factors regarding their importance for classification. The result has been shown how data mining models can help doctors in Libyan hospitals better understand cancer risk factors in order to make an accurate diagnosis. The research provides useful inputs for health decision-making bodies in Libya and beyond.

References:

Abdalkader. H, Hayajneh. Ferial. (2008).Effect of Night Shift on Nurses Working in Intensive Care Unitsat Jordan University Hospital. *European Journal of Scientific Research.* 23 (1), pp.70-86.

Adebamowo , C. A., *et al.* (2003). Waist-hip ratio and breast cancer risk in urbanized Nigerian women. *Breast Cancer Research,* 5(2), R18-R24.

American Cancer Society(2002).breast Cancer .Atlanta, GA:American Cancer Society. Available at **http://www.cancer.org/acs/groups/cid/documents/webcontent/003090-pdf.pdf.** 20-12-2007.

American Cancer Society. (2008) Guidelines for the early detection of cancer.www.cancer.org/docroot/ped/content/ped_2_3x_acs_cancer_detection_guidelines. Accessed on 15/1/2009.

Ames, B. N., Gold, L. S. and Willett, W. C. (1995). The causes and prevention of cancer. *Proceedings of the National Academy of Sciences of the United States of America,* 92(12), 5258-5265.

Amir, H. *et al.* (1996). Carcinoma of the male breast: A sexually transmitted disease? *East African Medical Journal,* 73(3), 187-190.

Apter, D., Reinila, M. and Vihko, R. (1989). Some endocrine characteristics of early menarche, a risk factor for breast cancer, are preserved into adulthood. *International Journal of Cancer,* 44(5), 783-787.

Awodele, O., Adeyomoye, A.A., Awodele, D.F., Fayankinnu, V.B., Dolapo, D.C. (2011).Cancer distribution pattern in southwestern Nigeria. Tanzania J. Health Res., 13(2),106-108.

Baker, J. A. *et al.* (2006). Consumption of coffee, but not black tea, is associated with decreased risk of premenopausal breast cancer. *Journal of Nutrition,* 136(1), 166-171.

Bernstein, Leslie and ROSS, R. K. (1993). Endogenous hormones and breast cancer risk. *Epidemiologic Reviews,* 15(1), 48-65.

Bernstein, Leslie *et al.* (2005). Lifetime recreational exercise activity and breast cancer risk among black women and white women. *Journal of the National Cancer Institute,* 97(22), 1671-1679.

Bernstein, Leslie *et al.,* (1994). Physical exercise and reduced risk of breast cancer in young women. *Journal of the National Cancer Institute,* 86(18), 1403-1408.

Biganzolil, E. *et al.,* (2003). Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors. *Annals of Oncology,* 14(10), 1484-1493.

Bishop, C.M, (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford, United Kingdom.

Black, T. R. (2002). Understanding social research. (2nd ed.). London: Sage.

Bray, Freddie, Mccarron, Peter and Parkin, Maxwell D. (2004). The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Research*, **6**(6), 229-239.

Breiman, L. (1984). Classification and regression trees. *Belmont, Calif.: Wadsworth.*

Brind JL, Chinchilli VM. Letter(2000) Abortion and breast cancer. J Epidemiol Community Health .**56**, 237-238.

Brinton, Louise A. *et al.*(1988). Menstrual factors and risk of breast cancer. *Cancer Investigation*, **6**(3), 245-254.

Brinton L, Daling J, Liff J,(1995). Oral contraceptives and breast cancer risk among younger women. J Natl Cancer Inst,87:827-835

Bryman, A. (2008). Social research methods. 3rd ed., *Oxford, Oxford University Press.*

Bunker, J. P., Houghton, J. and Baum, M. (1998). Putting the risk of breast cancer in perspective. *British Medical Journal*, **317**(7168), 1307-1309.

Calderon GAL *et al.* (2000) Risk factors of breast cancer in Mexican women. *Salud publicade Mexico*, **42**(1), 26–33.

Chan, Cheryl and MOUSAVI, Parvin (2005). Discovery of gene expression patterns across multiple cancer types. In: *Fifth IEEE Symposium on Bioinformatics and Bioengineering, 2005. Minneapolis, Minnesota, 19-21 October 2005.* IEEE, 121-128.

Chen, G., Warren, J. and Evans, J. (2008). Automatically generated consumer health metadata using semantic spaces. In: WARREN, J. et al. *Health Data and Knowledge Management: Proceedings of the Second Australasian Workshop on Health Data and Knowledge Management - Volume 80. Wollongong, NSW, January 2008.* Sydney: Australian Computer Society Inc., 9-16.

Chi, C., Street, N. and Wolberg, W. (2007). Application of artificial neural network-based survival analysis on two breast cancer datasets. In: *American Medical Informatics Association Annual Symposium 2007 : Biomedical and Health Informatics: from Foundations to Applications to Policy. Chicago, Illinois, 10-14 November 2007.* Bethesda, ML: American Medical Informatics Association, 130-134.

Cho, E., *et al.* (2006). Red meat intake and risk of breast cancer among premenopausal women. *Archives of Internal Medicine*, **166**(20), 2253-2259.

Colditz, G. A., *et al.* (2003). Physical activity and risk of breast cancer in premenopausal women. *British Journal of Cancer*, **89**(5), 847-851.

Collaborative Group on Hormonal Factors and Breast Cancer: (2002) Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. Lancet, 360:187-195.

Collaborative Group on Hormonal Factors in Breast Cancer (2001). Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological studies

including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet*, 358(9291), 1389-1399.

Committee on Gynecologic Practice. ACOG Committee Opinion(2009). Induced abortion and breast cancer risk. Obstetrics and Gynecology, 113(6): 1417–1418.

Corrado Gini (1912), I fattori demografici deirevoluzione delle nazioni, *Torino, Bocca.*

Dano, H. *et al.,* (2003). Socioeconomic status and breast cancer in Denmark. *International Journal of Epidemiology,* 32(2), 218-224.

De gonzalez, A. B. and Darby, S. (2004). Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. *The Lancet,* 363(9406), 345-351.

DecarliE, A. *et al.,* (1996). Age at any birth and breast cancer in Italy. *International Journal of Cancer,* 67(2), 187-189.

Eaker, S. *et al.,* (2009). Social differences in breast cancer survival in relation to patient management within a National Health Care System (Sweden). *International Journal of Cancer,* 124(1), 180-187.

Eaker, Sonja *et al.,* (2009). Breast cancer in the Thai Cohort Study: an exploratory case-control analysis. *The Breast,* 18(5-3), 299-303.

Ebrahimi, M., Vahdanini, M. and Montazeri, A. (2002). Risk factors for breast cancer in Iran: a case-control study. *Breast Cancer Research,* 4(5), R10.

El Mistiri M, Verdecchia A, Rashid I, *et al.* (2007).Cancer incidence in eastern Libya: The first report from the Benghazi Cancer Registry, 2003. *Int J Cancer 2007.*120(2), **392-397.**

Engeset, D. *et al.,* (2006). Fish consumption and breast cancer risk: the European Prospective Investigation into Cancer and Nutrition (EPIC). *International Journal of Cancer,* 119(1), 175-182.

Ferlay J, Bray F, Pisani P, Parkin DM, (2001). Globocan 2000: Cancer Incidence, Mortality and Prevalence Worldwide, *Version 1.0. IARC CancerBase No. 5. IARCPress,Lyon.*

Fioretti *et al.,* (2000). Menopause and risk of non-fatal acute myocardial infarction: an Italian case-control study and a review of the literature. *Human Reproduction,* 15(3), 599-603.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 1,* 179-188.

Floyd, C. *et al.* (1994). Prediction of breast cancer malignancy using an artificial neural network. *Cancer,* 74(11) 2944-2948.

Fogel, D., Wasson, E. and Boughton, E. (1995). Evolving neural networks for detecting breast cancer. *Cancer Letters,* 96(1), 49-53.

Fregene. A , Newman. L. (2005). Breast cancer in sub-Saharan Africa: how does it relate to breast cancer in African-American women? *Cancer,* 103(8) 1540-1550.

Furberg, H. *et al.* (2002). Environmental factors in relation to breast cancer characterised by p53 protein expression. *Cancer Epidemiology, Biomarkers and Prevention,* 11(9), 829-835.

Furundzic, D., Djordjevic, M. and Bekic, A. (1998). Neural networks approach to early breast cancer detection. *Systems Architecture,* 44(8), 617-633.

Gilliand, F. *et al.* (2001). Physical activity and breast cancer risk in Hispanic and non-Hispanic white women, *American Journal of Epidemiology,* 154(5), 442-450.

Giordano, S. H., Buzdar, A. U. and Hortobagyi, G. N. (2003). Breast cancer in men. *Annals of Internal Medicine,* 139(4), 305-305.

Golan, A. (2002). Information and entropy econometrics - An editor's view. *Journal of Econometrics,* 107(1-2), 1-15.

Gomez- -Ruiz, J. *et al.,* (2004) A neural network based model for prognosis of early breast cancer. *Applied Intelligence,* 20(3), 231-238.

Gonzaga, M. A. (2010). How accurate is ultrasound in evaluating palpable breast masses? *Pan,* 7(1), 1063-7788.

Graham, A. Colditz, Bernard A. Rosner, Frank E. Speizer(1996). Risk Factors for Breast Cancer According to Family History of Breast Cancer. *Oxford Journals Medicine* . 88(6), 365-371.

Grosan, C *et al.,* (2006). Evolving NNs for pharmaceutical research. In: *ICHIT'06: International Conference on Hybrid Information Technology, vol. 1. Jeju Island, Korea, 9-11 November 2006.* Piscataway, N.J.: IEEE, 13-19.

Gruvberger, S. *et al.* (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Research,* 61(16), 5979-5984.

Hamad HMA (2006). Cancer initiatives in Sudan. *Ann Oncol,11* (8), 32-6.

Hanf, V. and Gonder, U. (2005). Nutrition and primary prevention of breast cancer: foods, nutrients and breast cancer risk. *European Journal of Obstetrics and Gynecology,* 123(2), 139-149.

Hopfield, J. (1982) NNs and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences of the United States of America,* 79(8), 2554-2558.

Hsieh C. C. *et al.,* (1990). Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: associations and interactions in an international case-control study. *International Journal of Cancer,* 46(5), 796-800.

Hsieh, C. C. *et al.,* (1996). Does age at the last birth affect breast cancer risk? European Journal of Cancer, 32(1), 118-121.

Hu, M. B. *et al.,* (2008). Properties of wealth distribution in multi-agent systems of a complex network. *PhysicaA: Statistical Mechanics and its Applications,* 387(23), 5862-5867.

Huiyan, M. *et al.* (2006) Hormone-related risk factors for breast cancer in women under age 50 years by estrogen and progesterone receptor status: results from a case-control and a case-case comparison, [online]. *Breast Cancer Research,* 8(4), R39. Article from BioMed Central last accessed 22/03/2009 at: http://www.biomedcentral.com/.

Hunter, D. J., Willett, W. C. (1993). Diet, body size, and breast cancer. Epidemiologic Reviews, 15(1), 110-132.

Hussein, A. (2002). An evolutionary artificial neural networks approach for breast cancer detection. *Artificial Intelligence in Medicine,* 25 (3) 265-281.

JemalL, A. *et al.,* (2002). Cancer statistics, 2002. Ca: A Cancer Journal for Clinicians, 52(1), 23-4

Jerez-Aragones Jm, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz-Perez J, Alba-Conejo. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine,* 27(1), 45-63

Jemstrom, H. *et al.,* (2004). Breast-feeding and the risk of breast cancer in BRCA1 and BRCA2 mutation carriers. *Journal of the National Cancer Institute,* 96(14), 1094-1098.

John, E. M., Horn -Ross, P. L. and KOO, J. (2003). Lifetime physical activity and breast cancer risk in a multiethnic population the San Francisco bay area breast cancer study. *Cancer Epidemiology Biomarkers & Prevention,* 12(11), 1143-1152.

Kannel Wb, D'Agostino Rb, Cobb IL(1996). Effect of weight on cardiovascular disease. *Am J Clin Nutr* 63,419-422.

Katapodi, M. C. and Aouizerat, B. E. (2005). Do women in the community recognize hereditary and sporadic breast cancer risk factors? *Oncology Nursing Forum,* 32(3) 617-623.

Kelsey, J. L. and Berkowitz, G. S. (1988). Breast cancer epidemiology. *Cancer Research,* 48(20), 5615-5623.

Kelsey, J. L., Gammon, M. D. and John, E. M. (1993). Reproductive factors and breast cancer. Epidemiologic Reviews, 15(1), 36-47.

Kerr, D.J., Milbum,H.A., Arbuthnott,J(2007). Building Sustainable Cancer Capacity in Africa: Prevention, Treatment and Palliation. London.

Klecka WR. (1980).Discriminant analysis. Beverly Hills, CA, London.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine,* 23(1), 89-109.

Kordylewski. H, Graupe, D. and Liu, K. (2001). A novel large-memory neural network as an aid in medical diagnosis applications. *IEEE Transactions on Information Technology in Biomedicine,* 5(3), 202-209.

Krebs, E. E. *et al.,* (2006). Measures of adiposity and risk of breast cancer in older postmenopausal women. *Journal of the American Geriatrics Society,* 54(1), 63.

Kruger, W.M. and Apffelstaedt, J.P. (2007). Young breast cancer patients in the developing world: incidence, choice of surgical treatment and genetic factors. *South African Family Practice,* 49(9), 18-24.

La Vecchia, C., LEVI, F. and Lucchini, F. (1992). Descriptive epidemiology of male breast cancer in Europe. *International Journal of Cancer,* 51(1), 62-66.

Leon, D. A. (1989). A prospective study of the independent effects of parity and age at first birth on breast cancer incidence in England and Wales. *International Journal of Cancer / Journal International du Cancer,* 43(6), 986-991.

Li, C. I., Stanford, J. L. and Daling, J. R. (2000). Anthropometric variables in relation to risk of breast cancer in middle-aged women. *International Journal of Epidemiology,* 29(2), 208-213.

Little, M. (2001). Comparison of lung tumour mortality risk in the Japanese A-bomb survivors and in the Colorado Plateau uranium miners: support for the ICRP lung model. *International Journal of Radiation Biology,* 78(3), 145-163.

Lodish, H., Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J,(1995).Molecular cell biology 3rd ed. New York: Scientific American Books. Chap. 3.

Maclin.PS, Dempsey J, Brooks J, *et al.(* 1991) Using neural networks to diagnose cancer. *J Med Syst.* 15, 11-19.

Macmahon, B. *et al.,* (1982). Age at menarche, urine estrogens and breast cancer risk. *International Journal of Cancer,* 30(4), 427-431.

Marchbanks, P., Mcdonald, J., Wilson, H., Folger, S., Mandel, M., Daling, J., Bernstein, L., Malone, K., Ursin, G., Strom, B., Norman, S., Wingo, P., Burkman, R., Berlin, J., Simon, M., Spirtas, R. and Weiss, L. (2002) Oral contraceptives and the risk of breast cancer; *The New England Journal of Medicine, 346 (26), 2025-2032.*

Margolis, K. L. *et al.,* (2005). Physical activity in different periods of life and the risk of breast cancer: the Norwegian-Swedish women's lifestyle and health cohort study. *Cancer Epidemiology Biomarkers and Prevention,* 14(1), 27-32.

Mariani, L. *et al.* (1997). Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast Cancer Research and Treatment,* 44(2), 167-178.

Mcculloch, W.S. and Pitts, W. (1943) A logical calculus of ideas immanent in neural activity, *Bulletin of Mathematical Biophysics,* 5, 115-133.

Mcpherson, K., Steel, C. and Dixon, J. (2000). Breast cancer - epidemiology, risk factors, and genetics. *British Medical Journal,* 321(7261), 624-628.

Mctierana, A. *et al.,* (2003). Recreational physical activity and the risk of breast cancer in postmenopausal women the women's health initiative cohort study. *Journal of the American Medical Association,* 290(10), 1331-1336.

Mechanic L. *et al.* (2007). Polymorphisms in nucleotide excision repair genes, smoking and breast cancer in African Americans and whites: a population-based case-control study. *Environmental Health Perspectives,* 114(11), A642.

Meeske, K. *et al.,* (2004). Impact of reproductive factors and lactation on breast carcinoma in situ risk. *International Journal of Cancer,* 110(1), 102-109.

Mertens, A. J. *et al.,* (2006). Physical activity and breast cancer incidence in middle-aged women: a prospective cohort study. *Breast Cancer Research and Treatment,* 97(2), 209-214.

Mesa ,H., Cruz-Ranrirez,N, Hemandez-Jimenez,R.(2009)Aceto-white temporal pattern classification using k-NN to identify precancerous cervical lesion in colposcopic images. *Computers in Biology and Medicine.* 39 (9), 778-784.

Michels, K. B. *et al.,* (2002). Coffee, tea, and caffeine consumption and breast cancer incidence in a cohort of Swedish women. *Annals of Epidemiology,* 12(1), 21-26.

Minsky, M. and Papert, S. (1969). Perceptrons: an introduction to computational geometry. Cambridge, Mass, MIT Press.

Missmer, S. A. *et al.,* (2002). Meat and dairy food consumption and breast cancer: A pooled analysis of cohort studies. *International Journal of Epidemiology,* 31(1), 78-85.

Moorman, P., Terry, P. (2004). Consumption of dairy products and the risk of breast cancer: a review of the literature. *American Journal of Clinical Nutrition,* 80(1), 5-14.

Nagata, C., Hu, Y. H. and Shimizu, H. (1995). Effects of menstrual and reproductive factors on the risk of breast cancer: meta-analysis of the case-control studies in Japan. *Cancer Science,* 86(10), 910-915.

Neda, R. (2002). Data Mining with Decision Trees in the Gene Logic Database: A Breast Cancer Study. *Masters Dissertation, University ofSkovde, Sweden.*

Nichols, H. B. *et al.,* (2005). Differences in breast cancer risk factors by tumor marker subtypes among premenopausal Vietnamese and Chinese women. *Cancer Epidemiology Biomarkers and Prevention,* 14(1), 41-47.

Norsa'Adah, B. *et al.* (2005). Risk factors of breast cancer in women in Kelantan, Malaysia. *Singapore Medical Journal,* 46(12), 698-705.

Ohno-Machado L. (1996). Medical Applications of Artificial Neural Networks: Connectionist Models of Survival .*Ph.D, thesis. Departments of Computer Science and Medicine. Stanford University.*

Okobia, M. *et al.* (2006). Case-control study of risk factors for breast cancer in Nigerian women. *International Journal of Cancer,* 119(9), 2179-2185.

Ozkan, M., Dawant, B. and Maciunas, R. (1993). Neural-network-based segmentation of multi-modal medical images: a comparative and prospective study. *IEEE Transactions on Medical Imaging,* 12(3), 534-544.

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global Cancer Statistics, 2002. *CA: A Cancer Journalfor Clinicians,* 55(2), 74-108.

Pendharka, P. *et al.* (1999). Association, statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Systems with Applications,* 17(3), 223-232.

Pharoah, P. D. P. *et al.* (1997). Family history and the risk of breast cancer: a systematic review and meta-analysis. *International Journal of Cancer,* 71(5), 800-809.

Phelps, H. M. and Phelps, C. E. (1988). Caffeine ingestion and breast cancer: a negative correlation. *Cancer,* 61(5), 1051-1054.

Radice, D. and Redaelli, A. (2003). Breast cancer management: quality-of-life and cost considerations. *PharmacoEconomics,* 21(6), 383-396.

Rautalahti, M. *et al.,* (1993). Lifetime menstrual activity - indicator of breast cancer risk. *European Journal of Epidemiology,* 9(1), 17-25

Razavi, A. (2007). Applications of knowledge discovery in quality registries - predicting recurrence of breast cancer and analyzing non-compliance with clinical guidelines. *Unpublished thesis (PhD), Linkoping University.*

Ripley, B. D. (1994). Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society B.* 56(3): 409^437.

Robert, *S. A. et al.,* (2004). Socioeconomic risk factors for breast cancer: distinguishing individual-and community-level effects. *Epidemiology,* 15(4), 442-450.

Romieu, I. *et al.,* (1996). Breast cancer and lactation history in Mexican women. *American Journal of Epidemiology,* 143(6), 543-552.

Ronco, A. (1999). Use of artificial neural networks in modelling associations of discriminant factors: towards an intelligent selective breast cancer screening. *Artificial Intelligence in Medicine,* 16(3), 299-309.

Rosenberg, L. *et al.,* (1985). Breast cancer and the consumption of coffee. *American Journal of Epidemiology,* 122(3), 391-399.

Rosenblatt, F. (1962). Principles of Neurodynamics: perceptrons and the theory of brain mechanisms. *Washington, Spartan Books.*

Rumelhart, D., Hinton, G. and Williams, R. (1986). Learning internal represen-tations by error propagation. In: McClelland, J., Rumelhart, D. and the Pdp Research Group (eds). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol.l: Foundations.* Cambridge, Mass. and London, MIT Press, 318-362.

Sargent, D. J. (2001). Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer,* 91(8 Supp.), 1636-1642.

Sasco, A. J., Lowenfels, A. B. and Jong, P. P. D. (1993). Review article: Epidemiology of male breast cancer: a meta-analysis of published case-control studies and discussion of selected aetiological factors. *International Journal of Cancer,* 53(4), 538-549.

Sattin, R. *et al.,* (1985). Family history and the risk of breast cancer. *Journal of the American Medical Association,* 253(13), 1908-1913.

Setiono, R. (1996). Extracting rules from pruned neural networks for breast cancer diagnosis. *Artificial Intelligence in Medicine,* 8(1), 37-51.

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine,* 18(3), 205-219.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal.* 27, 379-423.

Shantakumar S, Terry Mb, Teitelbaum SL, Britton JA, Millikan RC, Moorman PG, Neugut Al, Gammon MD (2007): Reproductive factors and breast cancer risk among older women. Breast Cancer Res Treat, 102:365-374.

Stephenson, G. D. and Rose, D. P. (2003). Breast cancer and obesity: an update. *Nutrition and Cancer,* 45(1), 1-16

Sun, C. L. *et al.,* (2006). Green tea, black tea and breast cancer risk: a meta-analysis of epidemiological studies. *Carcinogenesis,* 27(7), 1310-1315

Susan Jordana, Lynette Lim, Duangkae Vilainerun (2009).Breast cancer in the Thai Cohort Study: An exploratory case-control analysis. Breast, 18, 299-303.

Sweeney, C. *et al.,* (2004). Risk factors for breast cancer in elderly women. *American Journal of Epidemiology,* 160(9), 868-875

Talamini, R. *et al.,* (1996). The role of reproductive and menstrual factors in cancer of the breast before and after menopause. *European Journal of Cancer,* 32(2), 303-310.

Tan DS, Marchio C, Reis-Filho JS(2008). Hereditary breast cancer: from molecular pathology to tailored therapies. J Clin Pathol,61: 1073 - 82.

Tang, L., Kacprzynski, G. J., Goebel, K., & Vachtsevanos, G. (2009). Methodologies for Uncertainty Management in Prognostics. *IEEE Aerospace Conference, Big Sky, MT.*

Tavani, A. *et al.* (1999) Risk factors for breast cancer in women under 40 years. *European Journal of Cancer,* 35(9), 1361-1367.

Taylor, C. and Mwitondi, K. (2001) Robust methods in data mining - in spatial statistics? In: *Proceedings of the Leeds Annual Statistical Research Conference: July 2001,* Leeds University Press, 67-70.

Terry, P. *et al.,* (2002). Fish consumption and breast cancer risk. *Nutrition and Cancer,* 44(1), 1-6

Thongkam,J Guandong Xu, Yanchun Zhang, Fuchun Huang(2009).Toward breast cancer survivability prediction models through improving training space. *Victoria University, Australia, Expert Systems with Applications.* 36, 12200-12209.

Titus-Emstoff, L. *et al.,* (1998). Menstrual factors in relation to breast cancer risk. *Cancer Epidemiology Biomarkers and Prevention,* 7(9), 783-789.

Trapido, E. J. (1983). Age at first birth, parity, and breast cancer risk. *Cancer*, **51**(5), 946-948.

TrichopoulosS, D. *et al.*, (1983). Age at any birth and breast cancer risk. *International Journal of Cancer*, **31**(6), 701-704.

Ture, M., Tokatli, F. and Kurt, I. (2009). Using Kaplan-Meier analysis together with decision tree methods in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications: an International Journal*, **36**(2), 2017-2026.

Uhrhammer, Nancy et al., (2008). BRCA1 mutations in Algerian breast cancer patients: high frequency in young, sporadic cases *International Journal of Medical Sciences*, **5**(4), 97-202.

United Nations (2008). The Millennium Development Goals report 2008. New York: United Nations.

Ursin G, Ross RK, Sullivan-Halley J, *et al.* (1998) Use of oral contraceptives and risk of breast cancer in young women. Breast Cancer Res Treat.;50:175-184.

Van den Brandt, P. A. *et al.*, (2000). Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *American Journal of Epidemiology*, **152**(6), 514-527.

Vorobiof, Daniel A., Sitas, Freddy, Vorobiof, Gabriel (2001). Breast cancer incidence in South Africa. *Journal of Clinical Oncology*, **19**(18S), 125s-127s.

Warwick, J. *et al.* (2003). Breast density and breast cancer risk factors in a high-risk population. *Breast*, **12**(1), 10-16.

Webster, T. F. *et al.*, (2008). Community-and individual-level socioeconomic status and breast cancer risk: multilevel modeling on Cape Cod, Massachusetts. *Environmental Health Perspectives*, **116**(8), 1125-1129.

Wei, J. *et al.* (1998). Understanding artificial NNs and exploring their potential application for the practicing urologist. *Urology*, **52**(2), 161-172.

Wilding, P. *et al.* (1994). Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. *Cancer Letters*, **77**(2-3), 145–153.

Wilson, Pw, D'Agostino Rb, Levy D, *et al*(1998). Prediction of coronary heart disease using risk factor categories. *Circulation*. **97**:1837–1847.

Wingo, Phyllis A. *et al.*, (1997). The risk of breast cancer following spontaneous or induced abortion. *Cancer Causes and Control*, **8**(1), 93-108.

Wohlfahrt, J. and MELBYE, M. (2001). Age at any birth is associated with breast cancer risk. *Epidemiology*, **12**(1), 68-73.

World Health Organization (WHO. 2008). The impact of cancer - Nigeria. http://www.who.int/infobase/report.aspx.Accessed on 23/92008.

World Health Organization. 58th World Health Assembly Approved Resolution on Cancer Prevention and Control (2005)Geneva, Switzerland: World Health Organization.

Woten, D. (2006). Artificial NNs for breast cancer detection using micro antennas. *Unpublished thesis (M.Sc.), University of Arkansas.*

Xiong, X. *et al.,* (2005). Analysis of breast cancer using data mining and statistical techniques. In: *6th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/ Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks. Towson University, Towson, ML.* Los Alamitos, CA, IEEE, 82-7.

Yost, K. *et al.,* (2001). Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. *Cancer Causes and Control,* 12(8), 703-711.

Zheng, T. *et al.,* (2000). Lactation reduces breast cancer risk in Shandong province, China. *American Journal of Epidemiology,* 152(12), 1129-1135

Ziegler, R. G. *et al.,* (1996). Relative weight, weight change, height, and breast cancer risk in Asian-American women. *Journal of the National Cancer Institute,* 88(10), 650-660.

Zikmund, W. (1999) Business Research Methods. 6th ed., Fort Worth, Dryden, and London, Harcourt Brace.

Ziv, E. *et al* (2006). Genetic ancestry and risk factors for breast cancer among Latinas in the San Francisco Bay area. *Cancer Epidemiology, Biomarkers & Prevention,* 15(10), 1878-1885.

Zografos, G., Panou, M. and Panou, N. (2004). Common risk factors of breast and ovarian cancer: recent view. *International Journal of Gynecological Cancer,* 14(5), 721-740.

Zurada, J. (2007) Rule induction methods for credit scoring. *The Review of Business Information Systems,* 11(2), 11-21.

Bibliography:

Alex, M.J., Nixon, R.N.(2009) .Insilico Docking Studies on Anticancer Drugs for Breast Cancer. *Computer Science and Information Technology - Spring Conference, 2009. Singapore.* 567 – 570.

Ali, A., Tufail, A., Khan, U., and Kim, M. (2009). A survey of prediction models for breast cancer survivability. *In Proceedings of the 2nd International Conference on Interaction Sciences, Information Technology, Culture and Human*, 1259-1262.

Alonso, O., Massardo T, Delgado LB, Horvath J, Kabasakal L, Llams-Olier A et al, (2001). Is 99mTc-Sestamibi scintimammography complementary to conventional mammography for detecting breast cancer in patients with palpable masses? *The Journal of Nuclear Medicine,* **42**(11), 1614-1621.

Aronowitz, Robert A. (2007). *Unnatural history: Breast cancer and American society.* Cambridge, N.Y., Cambridge University Press.
Baev K. (1998) Biological neural networks: hierarchical concept of brain function. *Boston, Birkhäuser.*

Baker, J. A. *et al.* (1995) Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology*, **196**(3), 817-822.

Beral V (2002). Breast cancer and breastfeeding: Collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50,302 women with breast cancer and 96,973 women without the disease. *Lancet*, **360**, 187-195.

Blake, C. and Merz, C. (1998). UCI Machine Learning Repository. [online] *University of California, Irvine, Dept. of Information and Computer Sciences*. Last accessed 12/03/2009 at http://archive.ics.uci.edu/ml/.

Bonneterre, J. *et al.*, (2000) Anastrozole versus tamoxifen as first-line therapy for advanced breast cancer in 668 postmenopausal women: results of the Tamoxifen or Arimidex randomized group efficacy and tolerability study. *Journal of Clinical Oncology*, **18**(22), 3748-3757.

Burke, H. B., Rosen, D.B. and Goodman, P.H. (1994). Comparing artificial neural networks to other statistical methods for medical outcome prediction. In: IEEE World Conference on Computational Intelligence, 1994, Vol. 4. Orlando, FL, 27 June – 2 July 1994. [New York] IEEE Neural Networks Council; Piscataway, N.J., 2213 - 2216.

Carrasco, M. J. (2004). African American outreach resource manual. NAMI Multicultural Action Center.

Colditz, G. A., Rosener, B. A. and Speizer, F. E. (1996). Risk factors for breast cancer according to family history of breast cancer. *Journal of the National Cancer Institute*, **88**(6), 365-371.

Coyle, Y. M. (2004). The effect of environment on breast cancer risk. *Breast Cancer Research and Treatment*, **84**(3), 273-288.

Cruz-Ramirez, N., *et al.* (2009). Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks. *Applied Soft*

Computing Journal, 9(4), 1331-1342.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine,* 34(2), 113-127.

Disease Control Priorities Project (DCPP) (2007). Controlling cancer in developing countries; prevention and treatment strategies merit further study, *www.dcp₂.org*.

Duda, R. O., Hart P. E. and Stork, D. G. (2001). Pattern classification and scene analysis. 2nd ed., New York, Chichester, Wiley.

Efird, J. and Nielsen, S. (2008). A method to compute multiplicity corrected confidence intervals for odds ratios and other relative effect estimates. *International Journal of Environmental Research and Public Health,* 5(5), 394-398.

El Saghir, N. S. *et al.,* (2007). Trends in epidemiology and management of breast cancer in developing Arab countries: a literature and registry analysis. *International Journal of Surgery 5(4), 225-233 http://www.ncbi.nlm.nih.gov/pubmed/17660128 accessed 12/06/2010.*

Esopsito, F., Malerba, D. and Semeraro, G. (1993). Decision tree pruning as a search in the state space. In: *ECML-93: European Conference on Machine Learning. Vienna, Austria, April 5-7, 1993.* Berlin and London: Springer-Verlag, 165-184.

Fogel, D., Wasson, E. and Porto, V. (1997). A step toward computer-assisted mammography using evolutionary programming and neural networks. *Cancer Letters,* 119(1), 93-97.

Gammon, Marlie D., *et al.* (1995). Cigarette smoking and breast cancer risk among young women (United States). *Cancer Causes and Control,* 9(6), 583-590 http://www.springerlink.com/content/114126408610648k accessed 24/4/2010

Gayther, Simon A. *et al.,* (1995). Germiline mutations of the BRCA1 gen in breast and ovarian cancer families provide evidence for a genotype-phenotype. *Nature Genetics,* 11(4), 428-433.

Golan, A. (2006).Information and entropy econometrics: A review and synthesis. *Found. Trend. Econom,* 2(1-2), 1-145.

Green, J. *et al.,* (1997). Family communication and genetic counseling: the case of hereditary breast and ovarian cancer. *Journal of Genetic Counseling,* 6(1), 45-60.

Hall, I. J. *et al.,* (2005). Comparative analysis of breast cancer risk factors among African-American women and white women. *American Journal of Epidemiology,* 161(1), 40-51.

Harirchi, I. *et al.* (2000). Breast cancer in Iran: a review of 903 case records. *Public Health,* 114(2), 143-145.

Hawkins, D. M., Basak, S. C. and Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences,* 43(2), 579-586.

Holmberg, Mats Lambe (2000). Social differences in breast cancer survival in relation

to efficacy and tolerability study. *Journal of Clinical Oncology,* 18(22), 3748-3757.

Hsieh, F., Bloch, D. and Larsen, M. (1998) A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine,* 17(14), 1623-1634.

Kalache, A, Maguire, A. and Thompson, S.G. (1993). Age at last full-term pregnancy and risk of breast cancer. *The Lancet,* 341 (8836), 33-36.

Khan MU, Choi JP, Shin H, Kim M, 2008, "Predicting Breast Cancer Survivability using Fuzzy Decision Trees.

Kulldorff, M. *et al.,* (1997). Breast cancer clusters in the northeast United States: a geographic analysis. *American Journal of Epidemiology,* 146(2), 161-170.

Laurance, Jeremy. (2006). Breast cancer cases rise 80% since seventies. *The Independent, 29 September.*

Lavrac, N. (1999). Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine,* 16(1), 3-23.

Lee, S. and Zelen, M. (2006). A stochastic model for predicting the mortality of breast cancer. *Journal of the National Cancer Institute. Monographs,* (36), 79-86.

Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW, (2002). Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry,* 48(8), 1296-1304.

Lingwood, R. *et al.* (2008). The challenge of cancer control in Africa. *Nature Reviews Cancer,* 8(5), 398-403.

Llewellyn, C. D. *et al.,* (2004). An analysis of risk factors for oral cancer in young people: a case-control study. Oral Oncology, 40(3), 304-313.

M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Data Stream Mining," Data Mining and Knowledge Discovery Handbook, pp. 759-787, 2010.

Maskarinec, G. *et al.,* (2007). Ethnic and geographic differences in mammographic density and their association with breast cancer incidence. *Breast Cancer Research and Treatment,* 104(1), 47-56.

Mcallister, M. *et al.,* (1998). Men in breast cancer families: a preliminary qualitative study of awareness and experience. *British Medical Journal,* 35(9), 739-744.

McCullagh P, Nelder JA. (1983).Generalized Linear Models, *Chapman and Hall, London*

Mechanic L. *et al.* (2006). Polymorphisms in nucleotide excision repair genes, smoking and breast cancer in African Americans and whites: a population-based case-control study. *Carcinogenesis,* 27 (7), 1377-1385.

National Cancer Institute. SEER cancer statistics review, 1975-2003. Acute myeloid leukemia. Available from: *http://seer.cancer.gov/ statfacts/html/amyl.html?statfacts_page_amyl.html&x_l3&y_l8 (accessed 27/11/2007)*

Oviedo, S. (2004). Body size and breast cancer risk: findings from the European Prospective Investigation into Cancer and Nutrition (EPIC). *International Journal of Cancer,* **111**(5), 762-771.

Parkin, D. M. and Fernandez, L. M. G. (2006). Use of statistics to assess the global burden of breast cancer. *The Breast Journal*, **12**(Supp.), S70-S80.

Pharoah, P. D. P. *et al.*, (2008). Polygenes, risk prediction, and targeted prevention of breast cancer. *The New England Journal of Medicine*, 2796-2803.

Remennick, L., (2006).The challenge of early breast cancer detection among Immigrant and Minority Women in Multicultural Societies. *The Breast Journal*, **12**(1), S103-S110

Rohan, T., Mcmichael, A. and Baghurst, P. (1988). A population-based case-control study of diet and breast cancer in Australia. *American Journal of Epidemiology*, **128**(3), 478-489.

Setiono, R. and Huan. L. (1997). Neurolinear: from neural networks to oblique decision rules. *Neurocomputing*, **17**(1), 1–24.

Setiono, R. and Liu, H. (1995). Understanding neural networks via rule extraction. In: mellish, C. (ed) *Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Quebec, August 1995*. San Mateo, CA, Morgan Kaufmann, 480–487.

Sternfeld, B., Jacobs, M. K., Quesenberry Jr., C. P., Gold, E. B. and Sowers, M. (2002). Physical Activity and Menstrual Cycle Characteristics in Two Prospective Cohorts; American Journal of Epidemiology, Vol. 156, No. 5, pp 402-409.

Stewart, B. and Kleihues P. (eds.) (2003). World Cancer Report. Lyons, IARC Press and Oxford, Oxford University Press.

Trichopoulos, D., Macmahon, B. and Cole, P. (1972). Menopause and breast cancer risk. *Journal of the National Cancer Institute*, **48**(3), 605-613.

Trichopoulou, Antonia *et al.* (1995). Consumption of olive oil and specific food groups in relation to breast cancer risk in Greece. *Journal of the National Cancer Institute*, **87**(2), 110-116.

Tung, H. *et al.*, (1999). Risk factors for breast cancer in Japan, with special attention to anthropometric measurements and reproductive history. *Japanese Journal of Clinical Oncology*, **29**(3), 137-146.

Walker, A. R. P., Adam, F. I. and Walker, B. F. (2004). Breast cancer in black African women: a changing situation. *The Journal of the Royal Society for the Promotion of Health*, **124**(2), 81-85.

Whitley, E. and Ball, J. (2002). Statistics review 3: hypothesis testing and P values. *Critical Care*, **6**(3), 222-225.

Williams, C. K. O., Olopade, O. I. and Falkson, C. I. (eds.) (2006). Breast cancer in women of African descent. *Dordrecht, The Netherlands, Springer.*

Wolff, M. S. and Weston, A. (1997). Breast cancer risk and environmental exposures. *Environmental Health Perspectives*, 105(Supp.4), 891-896 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1470027/ accesses 03/08/2010.

World Health Organization (2009). *Fact sheet N°297: Cancer,* [online]. Last accessed 22/03/2010 at: http://www.who.int/mediacentre/factsheets/fs297/en/.

Xu, Hengyi *et al.,* (2009). Application of semiconductor quantum dots for breast cancer cell sensing *proceedings of the 2009 2nd International Conference on Biomedical Engineering and Informatics, BMEI2009. Tianjin, China, 27-30 May 2009.* Piscataway, N.J., IEEE, 1-5.

Ya-Qin, L., Cheng, W and Lu, Z. (2009). Decision tree based predictive models for breast cancer survivability on imbalanced data, [online]. In: *The 3rd International Conference on Bioinformatics and Biomedical Engineering, Beijing, June 2009.* IEEE last accessed 23/04/2009 at http://ieeexplore.ieee.org/Xplore/dvnhome.isp.

Zupan, B., LAVRAC, N. and Keravnou, E. (1998). Data mining techniques and applications in medicine. *Artificial Intelligence in Medicine,* 16(1), p 1-2.

Appendix (A) Publication paper

Predicting breast cancer using combined K-fold cross-validated decision tree models

Mohamed Salem *(msalem2@my.shu.ac.uk)* ***and*** *Dr* ***Kassim Mwitondi*** *(k.mwitondi@shu.ac.uk)*
Sheffield Hallam University, Faculty of Arts, Computing, Engineering and Sciences

Abstract

Different forms of cancer have been widely studied and documented in various studies across the world. However, there have not been many similar studies in the developing countries - particularly on the African continent (Parkin, *et al.,* 2005). This paper seeks to uncover the geo-demographic occurrence patterns of the disease by applying decision tree models to learn the underlying rules in the overall behaviour of breast cancer. The data, 3,057 observations on 29 variables, obtained from four cancer treatment centres in Libya (2004-2008) were interrogated using multiple K-fold cross-validated decision tree models. The results from the selected optimal models exhibit greater accuracy and reliability as compared to using conventional decision models. The proposed strategy is therefore strongly recommended for use as a predictive tool in health and clinical centres to help minimise high costs of pathological tests. It is expected that the findings from this paper will provide an input into comparative geo-ethnic studies of cancer and provide informed intervention guidelines in the prevention and cure of the disease not only in Libya but also in other parts of the world.

Keywords

Breast cancer, cross validation, data mining, decision trees, over-fitting and risk factors

Introduction and problem description

Cancer is a major cause of death worldwide, with breast cancer claiming more lives than any other form of cancer. The disease accounts for over one fifth of the annual estimated 4.7 million diagnoses of all female cancers and it is the second most common tumour, after lung cancer, in both sexes. Ferlay *et al.,* (2001) report that worldwide, more than one million new developed countries. However, although breast cancer rates are high in many western countries, deaths from the disease have been decreasing as a result of improved screening and treatment. It is projected that by 2020 there would be 15 million new cases of cancer every year with 70% occurring in developing countries. Implicitly, African countries will have over a million new cancer cases a year and bearing in mind that they are likely to be the least prepared to face the problem, most patients are likely to face low survival rates. Hence, there is a pressing continent-wide need for a thorough study of factors predisposing to cancer.

Although there has been much research into the risk factors associated with this form of the disease, only a limited number of studies have investigated Africa. Fighting the spread of life-threatening diseases and providing cost-effective remedies is a global priority. Thus, the paper is motivated by the geo-ethnic patterns of breast cancer, the lack of in-depth breast cancer studies across the African continent and the socio-cultural variations among African populations. The lack of in-depth breast cancer studies on the African continent is particularly crucial which, according to Parkin *et al.,* (2002), limits the flow of documented cancer statistics from the continent. Thus, the paper will explores the risk factors associated with breast cancer in order to gain a better understanding of the disease, develop plans and strategies on how to deal with the factors and be able to fight the disease effectively.

The paper's ultimate goal is to investigate the geo-ethnic patterns of the disease using predictive modelling techniques in order to try and answer the following question: What is the nature of emerging patterns in the development of breast cancer among Libyan women? Its key objectives are two-fold - to discover and to understand the predisposition factors for breast cancer in Libya in order to help the relevant health authorities make informed interventions in the prevention and cure of cancer and to provide a basis for a global comparative analysis of the breast cancer studies. The paper is organised into five main parts - introduction, background, methods, results and summary and discussions.

Background of cancer studies in Africa

The cause of breast cancer is not yet known but there are certain risk factors that are linked to the disease. According to Katapodi *et al.,* (2005) approximately a quarter of breast cancers affect women under the age of 50, half of the cases occur between the ages of 50 and 69 and the remaining quarter among women aged 70 years or older. Research associated with these risk factors shows that there may be a link between breast cancer and marital status. Krueger and Apfelstaedt (2007) view the rising rate of breast cancer within the lower socio-economic groups in Africa as a serious problem for the African continent. Breast cancer among younger women is more prevalent in Africa than on other continents. Particularly worrying is the fact that breast cancer mortality rates are already higher in Africa than in the richer regions of the world with improved access to efficiently delivered therapies.

Other data suggests that breastfeeding mothers are at low risk. The relationship between body size and risk of breast cancer has been the subject of numerous investigations. Being overweight or obese has also been reported to increase the risk of postmenopausal breast cancer in premenopausal women (van den Brandit *et al.,* 2000). Many studies have found family history to be strongly correlated to breast cancer (Zografos *et al.,* 2004). The relationship between diet and the risk of breast cancer is often seen in terms of specific foods with a strong positive relationship been reported between the disease and the consumption of red meat Zografos *et al.,* (2004). Both Zografos *et al.* (2004) and Mertens *et al.* (2006) have investigated the relationship between physical activity and this risk and found an inverse relationship between the two. Research has also shown an inverse relationship between breast cancer and breastfeeding. Risk factors associated with breast cancer due to the reproductive cycle are similar in different groups. Susan Jordan *et al.,* (2009) found no significant difference in the occurrence of breast cancer between women who had children and those who had none. Earlier studies revealed that early menarche and late menopause could be linked to breast cancer. There has also been some controversy regarding the relationship between oral contraception and breast cancer Marchbanks *et al.,* (2002).

Methods and implementation strategy

Apparently, the foregoing studies were carried out using a variety of statistical methods applied on data obtained from various sources which entails data and model validation. Taylor and Mwitondi (2001) show that the nature of accuracy and reliability of analyses of large data sets are typically dependent on data and the methods the combination of

which generates two types of randomness - randomness due to the sampled data and randomness due to the domain partitioning. Data were collected from four cancer centres on the Mediterranean coast for patients coming from as far south as Gat near the Algerian border and Al Qufra near the Sudanese and Chadian borders. The data sampled from the Libyan population of around five million people scattered over 1.8 million square kilometres exhibited a steady increase in the rate of breast cancer over the period of time, with the highest growth occurring in Benghazi and the lowest in Misurata as shown in Figure 0-1.

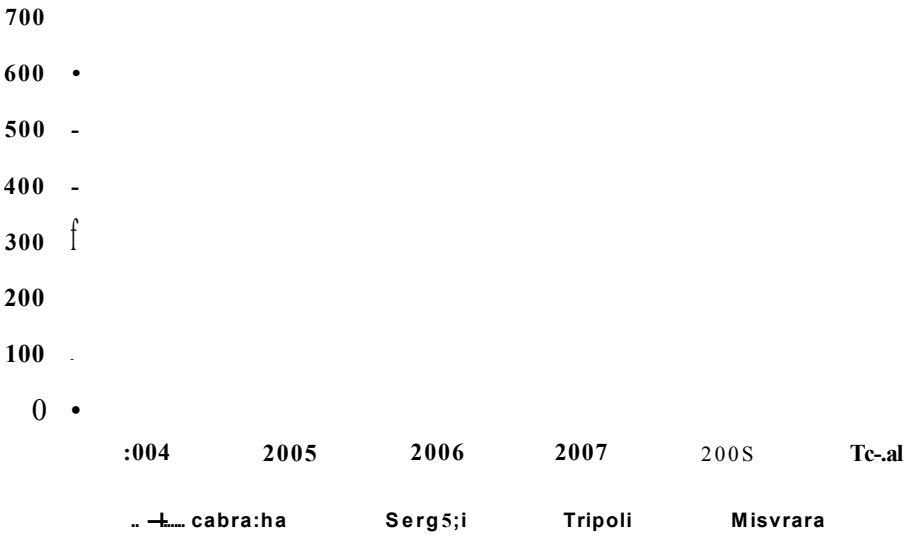


Figure 0-1: General trends of breast cancer in Libya (Constructed from sampled data)

To uncover breast cancer patterns, we apply a combination of classification tree models to learn the rules from the training data and use them to carry out predictions. The models sequentiallysplit thetraining dataset X into groups based onselected data variable thresholds.To split the data, the modelneeds someknowledge of eachof the classes.

$$x = \begin{matrix} \blacksquare\text{'11'}\ X\ 12^{*} & X\ |p \\ X\ 2\ 1\ '\ X\ 22\cdot \text{“^23} & ^{\wedge}29 \end{matrix}$$

$$X\ N\ |\ '\ X\ N\ 2\ '\ X\ N\ 3\ '\text{"}\ \text{'X}\ Np$$

The graphical representation of the data distribution is provided in Figure 0-2 in which the priors for developing or not developing breast cancer are given as n 1 and respectively. Then if the respective group distributional densities are -^W and the breast cancer prediction rule will allocate new cases to group 1 if

$$*\,i/i\,(*) \qquad >$$
$$X, fi(x) + x\,2f\,2(x) \quad x\ If\,l(x) + x\,2f\,2(x)$$

otherwise the case is allocated to group 2. The rule is associated with a total prediction error $£ \sim £\backslash + £\,2$ where both $£'$ and $£\,2$ depend on the priors and the densities. Our interest is therefore in minimising the rate of overlapping between the two groups.

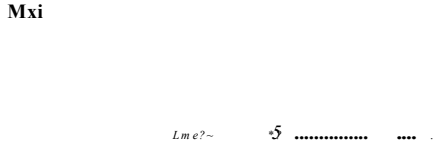


Figure 0-2: A typical bi-modal scenario for describing the presence or absence of breast cancer

Based on a binary scenario in which patients were allocated into one of the two groups - with and without breast cancer - multiple decision tree models were applied on the data matrix X. The splits are determined by the adopted a measure of impurity - typically examples include the Gini, Deviance, Entropy and the Chi-Square. The Gini index is computed as

$$pf = \mathbf{i} - \mathbf{Y}\ pf$$
$$\mathbf{i=1} \qquad \mathbf{i=l} \qquad \mathbf{i=l} \qquad \mathbf{2=1}$$

where Pi is the group's proportion and k isthe number of groups. The index measures the homogeneity or heterogeneity of the samples by examining the probability of two patients drawn at random belonging to the same group. One of the main issues in decision tree modelling is over-fitting which can be avoided by stopping growing the tree before it adapts to individual cases. One way of avoiding over-fitting is to cross-validate the model during its training - that is, estimating the expected level of fit of a model to a data set different to the set used to train it for which we need an appropriate strategy. Our strategy, graphically presented in Figure 0-3 involves two main steps. In the first step, all data attributes are divided into the categories - demographical, geographical, control, uncontrolled and health condition. The second step involves testing the models for their relevance in predicting breast cancer.

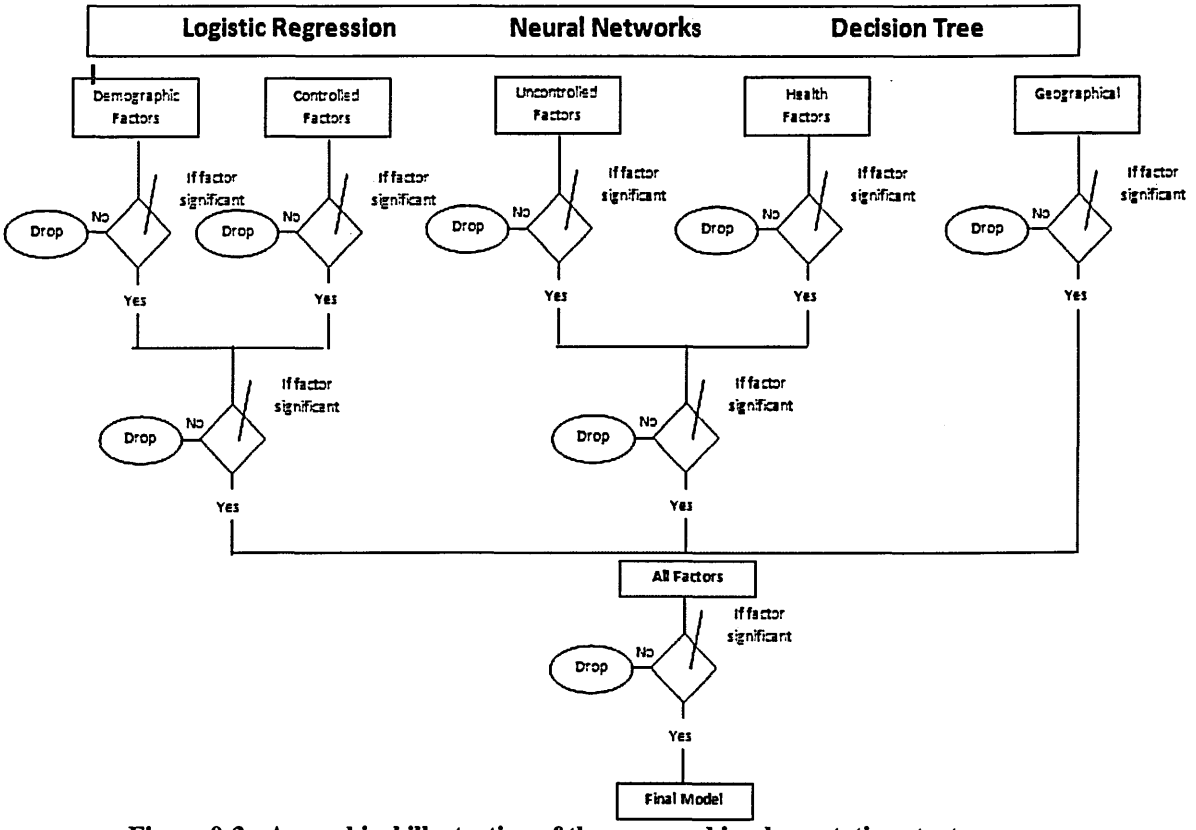
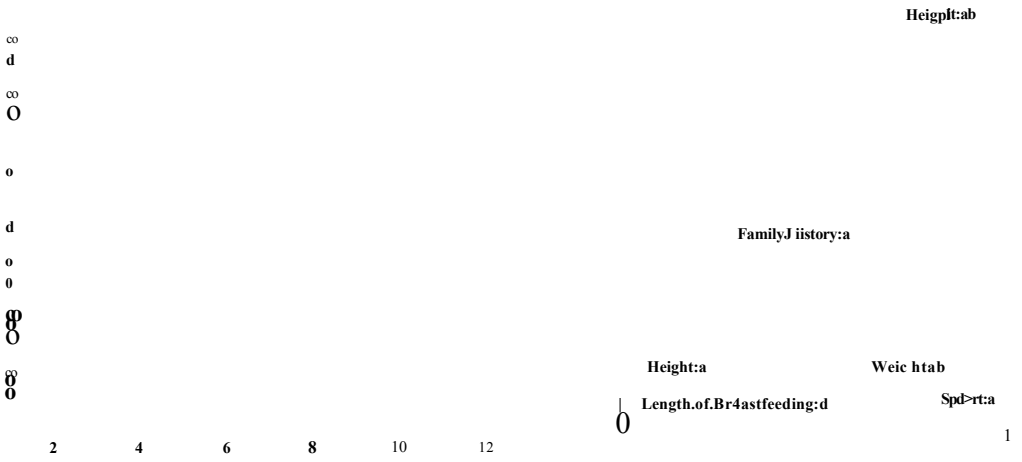


Figure 0-3 : A graphical illustration of the proposed implementation strategy

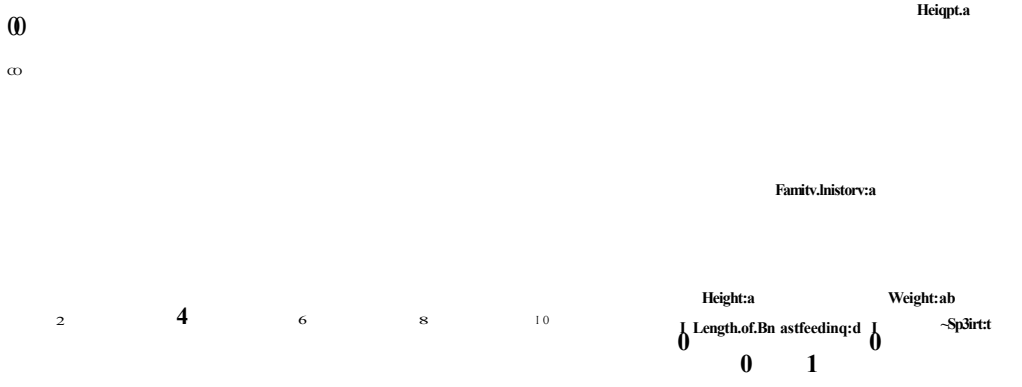
To test the models we apply the cross-validation method - randomly dividing the data into K-groups using one subset for validation and the remaining K-1 subsets for training the model. Each model is tested with respect to the data attributes' relevance to predicting breast cancer and the model is re-tested with the significant variables while the insignificant ones drop out until the final model is obtained. That is, For all risk factors V(i=1,2,...,p) the test error for each tree model is measured based on a risk factor at a time and only the model with the minimum classification error rate was considered for inclusion into the next iteration model based on Vi+1 risk factors. The process is repeated until there is no farther reduction in the misclassification error rate. Finally, the performances of the different decision trees are compared in order to determine the optimum model complexity for accuracy and reliability.

Results from three-fold cross-validation



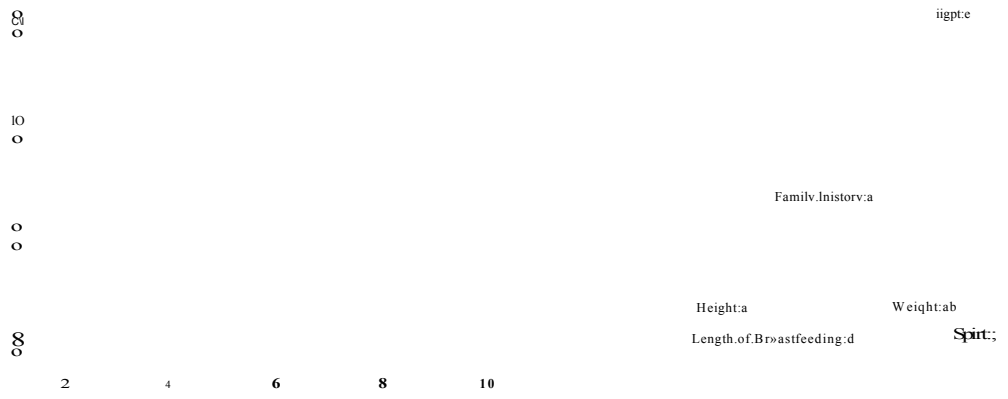
For the final model, the height, family history, weight, length of breastfeeding and participation in sporting activities emerged as the most important variables in discriminating the two groups.

Results from five-fold cross-validation



In this case, the final model yielded height, family history, weight, length of breastfeeding and participation in sporting activities as the most important variables.

Results from ten-fold cross-validation



Factors playing a role in the final model are height, family history, weight, Sport and length of breastfeeding.

Summary and discussions

	Three-fold		Five-fold		Ten-ifold	
Factors	Error rate	Tree size	Error rate	Tree size	Error rate	Tree size
Demographic	0.301	6	0.302	6	0.298	6
Control	0.119	7	0.119	5	0.123	5
Uncontrolled	0.093	6	0.093	6	0.093	6
Health	0.132	4	0.132	4	0.132	4
Demographic & control	0.116	7	0.124	5	0.124	5
Uncontrolled & Health	0.084	4	0.084	4	0.091	4
All factors	0.066	7	0.068	7	0.068	7

Table 1: Summary of results from the implementation of the adopted strategy

The combination of demographic and controlled factors does not yield markedly greater accuracy than controlled factors. However, some little improvement is returned by the uncontrolled and health factors, particularly for the three- and fivefold categories. From the final models consisting of the most important factors obtained by our plan, it can be observed that the error rate drops to a very satisfying 6.6 per cent. Overall, the decision tree classifier seems to be quite capable of allocate people into the appropriate classes.

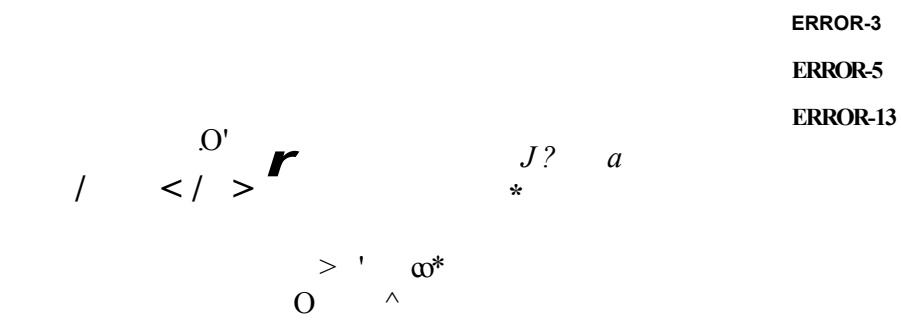


Figure 0-4 : three, five and ten-fold cross validation error patters (Source: authors' implementation)

The graphical version of the output in Table 1 is shown in Figure **0-4** in which it can be noted that isolating demographic factors as breast cancer predictors yields the highest error rates in all three validating cases. The near consistency of the predictive accuracy in all the remaining cases (under 15%) highlights the reliability of our adopted strategy. Finally, using our strategy to carry out analyses based on different combinations of cross-validation - i.e., validating on more than one set of training data revealed that the strategy with decision tree modelling was an appropriate method in predicting incidences of breast cancer. Again, there was no evidence of a geographical influence in the spread of the disease.

References

Ferlay, J., Bray, F., Pisani, P., Parkin DM. (2001). Globocan 2000: Cancer Incidence, Mortality and Prevalence Worldwide, *Version 1.0. IARC CancerBase No. 5. IARCPress,Lyon.*

Katapodi, M. and Aouizerat, B. (2005) Do women in the community recognize hereditary and sporadic breast cancer risk factors? *Oncology Nursing Forum,* **32** (3), 617-623.

Kruger, W. and Apffelstaedt, J. (2007) Young breast cancer patients in the world - incidence, choice of surgical treatment and genetic factors; South African Family Practice, **49** (9), 18-24.

Marchbanks, P., Mcdonald, J., Wilson, H., Folger, S., MandelL, M., Daling, J., Bernstein, L., Malone, K., Ursin, G., Strom, B., Norman S., Wingo, P., Burkman, R., Berlin, J., Simon, M., Spirtas, R. and Weiss, L. (2002) Oral contraceptives and the risk of breast cancer; *The New England Journal of Medicine, 346 (26), 2025-2032.*

Mertens, A., Sweeney, C., Shahar, E., Rosamond, W. and Folsom, A. (2006) Physical activity and breast cancer incidence in middle-aged women: a prospective cohort study; *Breast Cancer Research and Treatment Vol. 97, 209–214,* Springer.

Parkin D. M., Bray F., Ferlay, J., and Pisani P. (2005). Global Cancer Statistics, Cancer Journal for Clinicians, 55(2), 74-108.

Susan Jordana, Lynette Lim, Duangkae Vilainerun. (2009).Breast cancer in the Thai Cohort Study: An exploratory case-control analysis. Breast, 18, 299–303

Taylor, C. and Mwitondi, K. (2001) Robust methods in data mining – in spatial statistics? In: *Proceedings of the Leeds Annual Statistical Research Conference: July 2001*, Leeds University Press, 67-70.

Van den Brandt, P. A. *et al.*, (2000). Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. American Journal of Epidemiology, 152(6), 514-527

Zografos, G., Panou, M. and Panou, N. (2004). Common risk factors of breast and ovarian cancer: recent view. International Journal of Gynecological Cancer, 14(5), 721-740.

Appendix (B) Journey to Libya for Data Collection

Faculty of Art, Computing,

Sheffield Hallam University

Engineering and Sciences

Business Intelligence & Quantitative Modelling Research

Group, C3RI

Using Data Mining Techniques To Identify Breast Cancer Predictor Factors In Africa

Data Collection Report

From Libya

During 15-12-08 to 30 01-09

By

Mohamed Salem

February 09

CONTENT

JOURNEY TO LIBYA FOR DATA COLLECTION

1) **AFRICAN ONCOLOGY INSTITUTE SABRATHA**
2) **AL-JAMAHIRIYA HOSPITAL BENGHAZI**
3) **TRIPOLI CENTRAL HOSPITAL**
4) **NATIONAL CANCER INSTITUTE MISURATA**
5) **POSITION OF DATA COLLECTION CENTRES IN LIBYA**
6) **APPENDIX EVIDENCE TO PROVIDE DATA ENQUIRY.**

JOURNEY TO LIBYA FOR DATA COLLECTION

The aims of Journey visit four centres in Libya for Data Collection

1) African Oncology Institute Sabratha

The first visit to African Oncology Institute Sabratha on Monday 15-12-08.I meet Dr Hussein El Hashmi the Executive Director of the Institute and Dr Abugeila Abusaa the head of Cancer Registry in the Institute I explain to them the Aim and objects of my research how is importance to our continent after long discussion they express their willingness to provide us the Data enquiry, they following Day I meet the head of Statistical Department who Direct me to archives, there is no separate unit for breast cancer patient for that reason the first step we must separate the breast cancer folder before we start the data Collection after fourteen working day. The result summarized in Table (1) and presented in Figure (1).

Year	Number of all cancer patients	Number of breast cancer patients	Percent
2003	426	66	%15
2004	497	67	%13
2005	616	80	%13
2006	501	101	%20
2007	600	105	%18
2008	610	125	%20
Total	3250	544	

Table (1) the number of cancer patients in African Oncology Institute Sabratha.

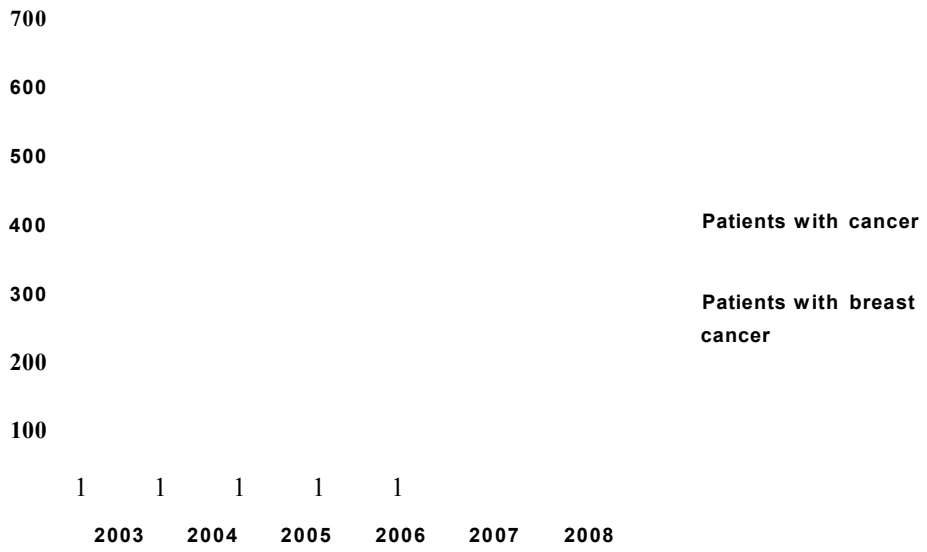


Figure (1) the number of cancer patients in African Oncology Institute Sabratha.

2) Al-Jamahiriya Hospital Benghazi

The first visit to Al-Jamahiriya Hospital Benghazi on Monday 29 -12-08 I meet Dr Mufid Elmistiri the head of Oncology Unit, I explain to him the aim of my visit , after long discussion we agreed to Stat by meeting Prof. A bdelfattah Zaied General Director of the Hospital on the second Day, on 30-12-08 meeting group with Dr Mufid Elmistiri, Dr Salem Aukhadra the head of internal Medicine Department and Prof. A bdelfattah, I explain to them the Aim and objects of my research how is importance to our country firstly and continent secondly the result meeting was the All agreed that is no objection for the Administration of the Hospital to carry out Statistical investigation of breast cancer in the Hospital of the Eastern Area.

They following Day I came with Dr Mufid Elmistiri to the archives, All folders patient mixed to gather no separate unit for breast cancer patient the same things happen in Sabratha the first step we must separate the breast cancer folder before we start the data Collection after fourteen working day. The result summarized in Table (2) and presented in Figure (2).

Year	Number of all cancer patients	Number of breast cancer patients	Percent
2004	410	95	%23
2005	493	112	%23
2006	569	113	%20
2007	706	148	%20
2008	705	173	%25
Total	2883	641	

Table (2) the number of cancer patients in Al-Jamahiriya Hospital Benghazi.

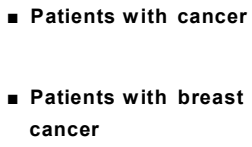


Figure (2) the number of cancer patients in Al-Jamahiriya Hospital Benghazi.

3) Tripoli Central Hospital

The first visit to Tripoli Central Hospital on 12 -01-09 I meet prof. Nuradin Aribi the head of General Surgery Department Tripoli Central Hospital, I introduce him myself and the aim of my visit and the Aim and objects of my research how is importance. The second Day meeting group with Prof. Aribi and the Team from unit Breast clinic; Dr Eman Hamad, Dr Thuraya A.Said and Aisha Zetoun, after a long discussion all of them they express their willingness to provide us the Data, the work seems easier than other centre because the unit of breast separate from other cancer unit. After Ten working days, the result summarized in Table (3) and presented in Figure (3).

Year	Number of breast cancer patients
2000	36
2001	35
2002	24
2003	40
2004	52
2005	62
2006	45
2007	72
2008	77
Total	443

Table (3) shown the number of breast cancer in Tripoli Central Hospital.

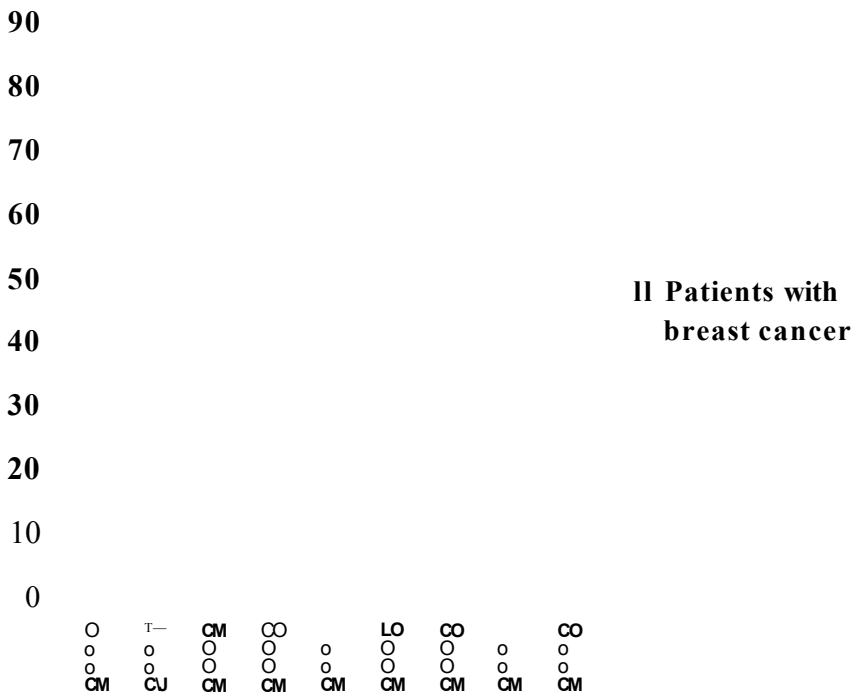


Figure (3) the number of cancer patients in Tripoli Central Hospital.

4) National Cancer Institute Misurata

The first visit to National Cancer Institute Misurata on 19 -01-09 I meet Dr Mohhammed Froka Chief of Medical Services direct me to Dr Abdulla Jebrel the head of medical Department, after discussion we agreed to start by meeting with Dr Mohamed Elfagieh the Director of Institute, the second Day we run the meeting at 10.00am.I explain to them the Aim and objects of my research how its importance to breast cancer patients firstly and our country and continent secondly. By the end of the meeting all agreed how the study is important and they accept to provide me the Data enquiry. After week working, the result summarized in Table (4) and presented in Figure (4).

Year	Number of breast cancer patients
2004	30
2005	8
2006	22
2007	20
2008	56
Total	136

Table (4) the number of cancer patients in National Cancer Institute Misurata.

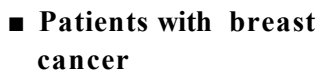


Figure (4) the number of cancer patients in National Cancer Institute Misurata.

In total from the four centres we get 1764 breast cancer patient

Position of data collection centres in Libya

Evidence to provide Data enquiry

J,V^IXii J ╪>WvTlVy

<Ul^M). £}ud J4*UI

AFRICAN ONCOLOCV INSTITUTE

few ln*iiitil Afrimiii ilr Onr«iloi*ii

SABRATHA - LIBYA

! i 'bij#

*L /'W Tf

To whom it may concern

We hove received a request of Mr. Moh&med salein from ShefTeild IlaIlain I nivcrsity to use the d*:U» in .statistieal department of the national Cancer instituie African Oncology institute.institute of African do oneoiogic Sabrata libya .
We would be delighted to give him a hand on that and wc are looking forward to seeing the very good results of his PhD study ,Inally we are ready to heljgMr \lohanwd for any things because he was vervgetfcf man
Fo: further assisiaticc plcsjxc free to contac; us.

/ X >T

Dr husekn amashmi.
Clief of th?mjlri«aii institUC

./ firA

(021) 3604053 j§i (024) 622326 ■622324 ■620963 - 620961

Evidence from African Oncology Institute Sabratha to provide Data enquiry

INTERNATIONAL OFFICE
K M JJbXiAl IRANS1 AT|; i *
A M) A! i i u i : \ llVi-tnON

JAMIL HAMED TATTER

•Mri r.p c.tc;: 45
JSJu. ata-l? Synhrc.crrr
fclei03i;<4i&9 -*e 3o.£4?2
*oM:0?IJ:,lSh04

tsiL sdl uaSUi

Ji.-fl3.il j i j s . l j l, } i . j j i 'V-:i. -V.'

1*It* (VSj W-jS 3
o' r.•'li?2vynh:>.*c:<4 ll•£/•% •**
3**T45 - i v u r. ^/3 4»a*94 .^s
op» j ai*
OWKIMMM

jVi4 -~t/jiifc*-JiUI• !'*2CJl'-A

Sheffield Hallom University
Sharpen your thinking

Uo whom it may concern (Assistance with firKnarrh Data Enquiry),

I wriio to introduce my l.lhU student. Mr. Mohamed Salem, who is working 011 Uw.ltopic: using data mining techniques In pinHict li pf.st cancer m-ti^»ihn^ factors ai Africa.

Wo will immensely appreciate every assistant o you could oiler to Mohamed widi regards to data collection (preliminary enquiries and collection), in nrvior to acilitate this research which is of stratn^ic impurteike to the continent.

I;t>r a background on diis work, Molujirrd is m Uk» second yea: ol tlie research, thu topic is fully approved by the Puralty Research Committee mid Hu: data and methodology stag's are now primary.

Thanks so much for your anticipated cooperation.

Yours sincerely,

Sheffield HaLuin University
Dr. Fati.ck EZE7UE (Diiortu; of Studies) (10/J2/20C8)

SeMMdi Load. JusiiK-ri? Inleliig&ncft <fe Vtodoli.-.g Kes&ueh (jrmip
Vifitrsg l-Yofewor of Stochastic Vtodclrngin Hnsnce *tr* Hiuwto^s Njbon.il.
Vlathematical Centre, Abuja, Nigeria

- **rijrm mrlya Ikwpital / "k'nghari**
- **Secretary of the >ept *ol* internal d »«Mse« / Cancers Unit**
- **No oVjeitien for the Administration ct the Hospital lit carry cut ti*» s.al:sl ical investigation .if :m*asl earner in tin* tmsplul dl" the eastern area.**

")i Salem Ahukhadra,
Head of Internal Medicine iVpt: No objection
^i.-iHtl. V)/12/3£W8

Signed In-.
Prof. Dv. Abddfattah Yvussef/j.J**
General DinvK* of the TTempitat
sofp/xtnn

IiLx#A-2?
ijTxInjjinj juiluxcJIihuk vS^nfijfiiiraj y fl_nk^ »s-f

Evidence from Al-Jamahiriya Hospital Benghazi to provide Data enquiry

^U?l

■WUJ Jvy

^yfdkaJl AiSl>£YI

j^j^jaJI 5dj!jh jmlulUu

/ : y KJmJI

/ } r J ------flj1l : .-.-jli

10+ku.iuuv 2009

IX-tii Di. Ptftrirk.

1 would Ii<e to inform that du PHD student Mr. Mohamed Salem Approached ino about data oolleUion am vrnir.g rarclnorn.j breast. Oiii Unit h/js 11'* objection for IIk Jati roiloclijn and vvr oiler him ossislaiKtf U> ^*<l n*iri*:1 data flhtut :*ur cancer jwiwis r^jjisttTed in Iho bicnst clinic, Ifipoli Central I kitipital l ripnli, Libya

Yours SJmrweJv

30tlb(JI'J JOUbUUI .

Evidence from Tripoli Central Hospital to provide Data enquiry

NATIONAL CACLR INSTITUTE

-r'UI *a » $]*JAail

,\

To *whom it may concern*

We have received a request of Mr. ***MohammedSalem /***

Sheffeild Hallant University **to use the data in statistical department of the National Cancc** Institute-misurata for a study involving breast cancer cases in Africa.**

We would be del ghtcd to give him a hand on that and we are looking forward to seeing the results of his study.

For further assistance please feel free to contact us.

Dt\ FROKA, Mohamrncd

Chii f of medical services

(4 . (

F-maihafd misura(a@Hocmaii.coin 051-2652751 051-2652719 • 051-2651645®

Evidence from National Cancer Institute Misurata

Appendix (C) Questionnaire

Risk factors for breast cancer

1 - Gender	Male				emale				
2 - Age	<20		20-30		31-40		41-50	51-60	61>
3 - Education Level	Illiterate			Primary		Secondary			High Education
4- Weight	<60		60-80		81-100			>100	
5- Height			≤150		151-170			>171	
6 - Employee	Yes				No				
7 - Does the work connected with radiation	Yes				No				
8 - socio-economic	<2000		2000-3600		>3600-6000		>6000		
9 - kind of Vegetable	Organic				Non-organic				
10- kind of meat	Red meat		Fish		Bird				
11- Sport	yes				No				
12 - Smoke	Yes				No				
13 - Drinking Alcohol	Yes				No				
14- Other diseases	yes				No				
15 -Inherited diseases	yes				No				
16- Marital State	Single			Married		Widowed/ Divorced			
17 - Family History of breast cancer	Yes				No				
18 - Age at menarche	≤12		13		14		≥15		

19 - Age at first Pregnancy	**<20**		**20-24**		**25-29**		**30-34**		**≥35**
20 - Breastfeeding	**Yes**				**No**				
21 -Length of Breastfeeding	**<1 year**		**1 year**				**≤2 year**		
22 - Number of Children	**<2**		**2-5**				**≥6**		
23 - effected after which birth									
24 - Spontaneous abortions	**yes**				**No**				
25- Age at last pregnancy	**≤29**		**30-34**		**35-39**		**≥40**		
26 - Age at menopause	**<45**		**≥45**						
27 - Oral contraceptive use	**yes**				**No**				
28 - Duration of oral contraceptive use	**<2**		**2-5**		**>5**				
29- Place of accommodation.	**West**		**East**		**South**		**Middle**		

Appendix (D) R CODES

R CODES

Logistic Regression Code

```
w=read.csv("fatmha.csv",sep=";",header=T)
library
Gender=as.factor(w$Gender)
Age=as.factor(w$Age)
Education.level=as.factor(w$Education.level)
Weight=as.factor(w$Weight)
Height=as.factor(w$Height)
Employee=as.factor(w$Employee)
work.connected.with.rediation=as.factor(w$work.connected.with.rediation)
Socio.Econmic=as.factor(w$Socio.Econmic)
Kind.of.Vegetable=as.factor(w$Kind.of.Vegetable)
Kind.of.Meat=as.factor(w$Kind.of.Meat)
Sport=as.factor(w$Sport)
Smoke=as.factor(w$Smoke)
Drinking.Alcohol=as.factor(w$Drinking.Alcohol)
Other.diseases=as.factor(w$Other.diseases)
Inherited.diseases=as.factor(w$Inherited.diseases)
Marital.State=as.factor(w$Marital.State)
Family.history=as.factor(w$Family.history)
Age.at.menarche=as.factor(w$Age.at.menarche)
Age.at.first.Pregnancy=as.factor(w$Age.at.first.Pregnancy)
Breastfeeding=as.factor(w$Breastfeeding)
Length.of.Breastfeeding=as.factor(w$Length.of.Breastfeeding)
Number.of.chlidren=as.factor(w$Number.of.chlidren)
Spontaneous.abortions=as.factor(w$Spontaneous.abortions)Age.at.last.pregnancy=as.fa
ctor(w$Age.at.last.pregnancy)
Age.at.menpause=as.factor(w$Age.at.menpause)
Oral.contraceptive.use=as.factor(w$Oral.contraceptive.use)
Duration.of.oral.contraceptive.use=as.factor(w$Duration.of.oral.contraceptive.use)
Place.of.Accommodation=as.factor(w$Place.of.Accommodation)
Breast.Cancer=as.factor(w$Breast.Cancer)
```

```
dat.dem=data.frame(Gender,Age,Education.level,Employee,Socio.Econmic,
Marital.State, Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numerie(0)
size=numerie(0)
V=l:(dim(dat.dem)[2]-1)
dem.names=names(dat.dem)
err.selct=as.vector(0)
R=0
L=numerie(0)
        for(N in l:length(V)){
mean.err=as.veetor(0)
Q=0
        for(v in V){ print(names(dat.dem)[v])
          all.err=numeric(0)
          R=c(L,v)
          for(k in 1:K){
dat.k=dat.dem[fold!=k,]
log=glm(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.dem)[2])],family=binomial)
   #### test error #####
pred.log=round(predict(log,dat.dem[fold==k,c(R,dim(dat.dem)[2])],
type="response"),0)
tab=table(dat.dem$Breast.Cancer[fold==k],pred.log)
err=l-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)   }
```

```
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=round(min(mean.err),3)
L=c(V[min.err],L)
V=V[-min.err] }
cbind(sort(err.selct),dem.names[L])
#########################################################################
############################## Control factors ###########################
dat.con=data.frame(Weight,Kind.of.Vegetable,Kind.of.Meat,Sport,Breastfeeding,
Length.of.Breastfeeding,Number.of.chlidren,Age.at.last.pregnancy,Oral.contraceptive.u
se,Duration.of.oral.contraceptive.use,Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))size=numeric(0)
size=numeric(0)
V=1:(dim(dat.con)[2]-1)
con.names=names(dat.con)
err.selct=as.vector(0)
R=0
L=numeric(0)
        for(N in 1:length(V)){
mean.err=as.vector(0)
Q=0
        for(v in V){ print(con.names[v])
          all.err=numeric(0)
          R=c(L,v)
          for(k in 1:K){
dat.k=dat.con[fold!=k,]
```

```
log=glm(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.con)[2])],family=binomial)
   #### test error #####
pred.log=round(predict(log,dat.con[fold==k,c(R,dim(dat.con)[2])], type="response"),0)
tab=table(dat.con$Breast.Cancer[fold==k],pred.log)
err=l-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)}
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=round(min(mean.err),3)
L=c(V[min.err],L)
V=V[-min.err]  }
cbind(sort(err.selct),con.names[L])
```

Control factors

```
dat.uncon=data.frame(Height,work.connected.with.rediation,Age.at.menarche,Age.at.fir
st.Pregnany,Age.at.menpause,Spontaneous.abortions,Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numerie(0)
V=l:(dim(dat.uncon)[2]-l)
uncon.names=names(dat.uncon)
err.selct=as.vector(O)
R=0
L=numeric(0)
 for(N in l:length(V)){
mean.err=as.vector(O)
```

```
Q=0
        for(v in V){ print(uncon.names[v])
         all.err=numeric(0)
         R=c(L,v)
         for(k in 1:K){
dat.k=dat.uncon[fold!=k,]
log=glm(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.uncon)[2])],family=binomial )
  #### test error #####
pred.log=round(predict(log,dat.uncon[fold==k,c(R,dim(dat.uncon)[2])],
type="response"),0)
tab=table(dat.uncon$Breast.Cancer[fold==k],pred.log)
err=1-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)}
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=round(min(mean.err),3)
L=c(V[min.err[1]],L)
V=V[-min.err[1]] }
cbind(sort(err.selct),uncon.names[L])
#########################################################################
############################## Health factors ############################
dat.health=data.frame(Other.diseases,Family.history,Inherited.diseases,Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numeric(0)
V=1:(dim(dat.health)[2]-1)
```

```
health.names=names(dat.health)
err.selct=as.vector(0)
R=0
L=numeric(0)
 for(N in 1:length(V)){
mean.err=as.vector(0)
Q=0
        for(v in V){ print(health.names[v])
         all.err=numeric(0)
         R=c(L,v)
         for(k in 1:K){
dat.k=dat.health[fold!=k,]
log=glm(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.health)[2])],family=binomial )
   #### test error #####
pred.log=round(predict(log,dat.health[fold==k,c(R,dim(dat.health)[2])],
type="response"),0)
tab=table(dat.health$Breast.Cancer[fold==k],pred.log)
err=1-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)  }
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=round(min(mean.err),3)
L=c(V[min.err[1]],L)
V=V[-min.err[1]] }
cbind(sort(err.selct),health.names[L])
#########################################################################
########## ####################Demo and Cont ##########################
dem.con=data.frame(Socio.Econmic,Education.level,Age,Employee,Weight,Sport,Leng
th.of.Breastfeeding,Kind.of.Meat,Breastfeeding,Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
```

```
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numeric(0)
V=l:(dim(dem.con)[2]-l)
dem.con.names=names(dem.con)
err.selct=as.vector(O)
R=0
L=numeric(0)
 for(N in l:length(V)){
mean.err=as.vector(0)
Q=0
          for(v in V){ print(dem.con.names[v])
            all.err=numeric(0)
            R=c(L,v)
            for(k in 1:K){
dat.k=dem.con[fold!=k,]
log=glm(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dem.con)[2])],family=binomial)
   #### test error #####
pred.log=round(predict(log,dem.con[fold==k,c(R,dim(dem.con)[2])],
type="response"),0)
tab=table(dem.con$Breast.Cancer[fold==k],pred.log)
err=l-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)  }
Q=Q+l
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=min(mean.err)
L=c(V[min.err[ 1]],L)
V=V[-min.err[l]]  }
cbind(sort(err.selct),dem.con.names[L])
```

```
data.frame(Height,work.connected.with.rediation,Spontaneous.abortions,Family.history,
Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+l
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numeric(0)
V=l:(dim(uncon.hlth)[2]-l)
uncon.hlth.names=names(uncon.hlth)
err.selct=as.vector(O)
R=0
L=numeric(0)
 for(N in l:length(V)){
mean.err=as.vector(O)
Q=0
          for(v in V){ print(uncon.hlth.names[v])
            all.err=numeric(0)
            R=c(L,v)
            for(k in 1:K){
dat.k=uncon.hlth[fold!=k,]
log=glm(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(uncon.hlth)[2])],family=binomial)
   #### test error #####
pred.log=round(predict(log,uncon.hlth[fold==k,c(R,dim(uncon.hlth)[2])],
type="response"),0)
tab=table(uncon.hlth$Breast.Cancer[fold==k],pred.log)
err= l-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)  }
```

```
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=min(mean.err)
L=c(V[min.err[1]],L)
V=V[-min.err[1]] }
cbind(sort(err.selct),uncon.hlth.names[L])
#########################################################################
############### #################All factors ################################
data.frame(Weight,Sport,Length.of.Breastfeeding,Kind.of.Meat,Breastfeeding,Height,F
amily.history,work.connected.with.rediation, Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numeric(0)
V=1:(dim(all)[2]-1)
all.names=names(all)
err.selct=as.vector(0)
R=0
L=numeric(0)
 for(N in 1:length(V)){
mean.err=as.vector(0)
Q=0
        for(v in V){ print(all.names[v])
          all.err=numeric(0)
          R=c(L,v)
          for(k in 1:K){
dat.k=all[fold!=k,]
```

```
log=glm(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(all)[2])],family=binomial )
  #### test error #####
pred.log=round(predict(log,all[fold==k,c(R,dim(all)[2])], type="response"),0)
tab=table(all$Breast.Cancer[fold==k],pred.log)
err=1-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err) }
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=min(mean.err)
L=c(V[min.err[1]],L)
V=V[-min.err[1]] }
cbind(sort(err.selct),all.name
```

Neural Networks Code

```
library(nnet)
w=read.csv("fatmha.csv",sep=";",header=T)
library(nnet)
Gender=as.factor(w$Gender)
Age=as.factor(w$Age)
Education.level=as.factor(w$Education.level)
Weight=as.factor(w$Weight)
Height=as.factor(w$Height)
Employee=as.factor(w$Employee)
work.connected.with.rediation=as.factor(w$work.connected.with.rediation)
Socio.Econmic=as.factor(w$Socio.Econmic)
Kind.of.Vegetable=as.factor(w$Kind.of.Vegetable)
Kind.of.Meat=as.factor(w$Kind.of.Meat)
Sport=as.factor(w$Sport)
Smoke=as.factor(w$Smoke)
Drinking.Alcohol=as.factor(w$Drinking.Alcohol)
Other.diseases=as.factor(w$Other.diseases)
Inherited.diseases=as.factor(w$Inherited.diseases)
Marital.State=as.factor(w$Marital.State)
Family.history=as.factor(w$Family.history)
Age.at.menarche=as.factor(w$Age.at.menarche)
Age.at.first.Pregnancy=as.factor(w$Age.at.first.Pregnancy)
Breastfeeding=as.factor(w$Breastfeeding)
Length.of.Breastfeeding=as.factor(w$Length.of.Breastfeeding)
Number.of.chlidren=as.factor(w$Number.of.chlidren)
Spontaneous.abortions=as.factor(w$Spontaneous.abortions)
Age.at.last.pregnancy=as.factor(w$Age.at.last.pregnancy)
Age.at.menpause=as.factor(w$Age.at.menpause)
Oral.contraceptive.use=as.factor(w$Oral.contraceptive.use)
Duration.of.oral.contraceptive.use=as.factor(w$Duration.of.oral.contraceptive.use)
Place.of.Accommodation=as.factor(w$Place.of.Accommodation)
Breast.Cancer=as.factor(w$Breast.Cancer)
```

```
dat.dem=data.frame(Gender,Age,Education.level,Employee,Socio.Econmic,
Marital.State, Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numeric(0)
V=l:(dim(dat.dem)[2]-1)
dem.names=names(dat.dem)
err.selct=as.vector(O)
R=0
L=numeric(0)
        for(N in l:length(V)){
mean.err=as.vector(0)
Q=0
        for(v in V){ print(names(dat.dem)[v])
          all.err=numeric(0)
          R=c(L,v)
          for(k in 1:K){
dat.k=dat.dem[fold!=k,]
net=nnet(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.dem)[2])],size = 20, rang = 0.1,
        decay = 5e-4, maxit = 200)
   #### test error #####
pred.net=predict(net,dat.dem[fold==k,c(R,dim(dat.dem)f2])],type="class")
tab=table(dat.dem$Breast.Cancer[fold==k],pred.net)
err=l-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)}
```

```
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=min(mean.err)
L=c(V[min.err],L)
V=V[-min.err] }
cbind(sort(err.selct),dem.names[L])
#########################################################################
############################## Control factors#############################
dat.con=data.frame(Weight,Kind.of.Vegetable,Kind.of.Meat,Sport,Breastfeeding,Lengt
h.of.Breastfeeding,Number.of.chlidren,Age.at.last.pregnancy,Oral.contraceptive.use,Du
ration.of.oral.contraceptive.use,Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numeric(0)
V=1:(dim(dat.con)[2]-1)
con.names=names(dat.con)
err.selct=as.vector(0)
R=0
L=numeric(0)
for(N in 1:length(V)){
mean.err=as.vector(0)
Q=0
 for(v in V){ print(con.names[v])
all.err=numeric(0)
        R=c(L,v)
        for(k in 1:K){
```

```
dat.k=dat.con[fold!=k,]
net=nnet(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.con)[2])],size _ ^    rang = Q.1,
        decay = 5e-4, maxit = 200)
  #### test error #####
pred.net=predict(net,dat.con[fold==k,c(R,dim(dat.con)[2])],type="class")
tab=table(dat.con$Breast.Cancer[fold==k],pred.net)
err=l-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err) }
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=min(mean.err)
L=c(V[min.err],L)
V=V[-min.err] }
cbind(sort(err.selct),con.names[L])
```

Control factors

```
dat.uncon=data.frame(Height,work.connected.with.rediation,Age.at.menarche,Age.at.fir
st.Pregnancy,Age.at.menpause,Spontaneous,abortions,Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
size=numeric(0)
V=1 :(dim(dat.uncon)[21-1)
uncon.names=names(dat.uncon)
err.selct=as.vector(O)
R=0
L=numeric(0)
```

```
 for(N in 1:length(V)){mean.err=as.vector(0)
Q=0
for(v in V){ print(uncon.names[v])
all.err=numeric(0)
 R=c(L,v)
 for(k in 1:K){
dat.k=dat.uncon[fold!=k,]
net=nnet(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.uncon)[2])],size = 30, rang =
0.1,
        decay = 5e-4, maxit = 200)
   #### test error #####
pred.net=predict(net,dat.uncon[fold==k,c(R,dim(dat.uncon)[2])],type="class")
tab=table(dat.uncon$Breast.Cancer[fold==k],pred.net)
err=1-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)}
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=min(mean.err)
L=c(V[min.err],L)
V=V[-min.err] }
cbind(sort(err.selct),uncon.names[L])
#########################################################################
############################## Health factors ############################
dat.health=data.frame(Other.diseases,Family.history,Inherited.diseases,Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(5)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
size=numeric(0)
```

```
size=numeric(0)
V=1:(dim(dat.health)[2]-1)
heath.names=names(dat.health)
err.selct=as.vector(0)
R=0
L=numeric(0)
 for(N in 1:length(V)){
mean.err=as.vector(0)
Q=0
for(v in V){ print(heath.names[v])
all.err=numeric(0)
R=c(L,v)
for(k in 1:K){
dat.k=dat.uncon[fold!=k,]
net=nnet(dat.k$Breast.Cancer~.,data=dat.k[,c(R,dim(dat.health)[2])],size = 30, rang =
0.1, decay = 5e-4, maxit = 200)
   #### test error #####
pred.net=predict(net,dat.health[fold==k,c(R,dim(dat.health)[2])],type="class")
tab=table(dat.health$Breast.Cancer[fold==k],pred.net)
err=1-sum(diag(as.matrix(tab)))/summary(fold)[k]
all.err=rbind(all.err,err)}
Q=Q+1
mean.err[Q]=mean(all.err)}
min.err=which(mean.err==min(mean.err))
err.selct[N]=min(mean.err)
L=c(V[min.err],L)
V=V[-min.err] }
cbind(sort(err.selct),uncon.names[L])
```

Decision Tree Code

```
w=read.csv("fatmha.csv",sep=";",header=T)
library(tree)
Gender=as.factor(w$Gender)
Age=as.factor(w$Age)
Education.level=as.factor(w$Education.level)
Weight=as.factor(w$Weight)
Height=as.factor(w$Height)
Employee=as.factor(w$Employee)
work.connected.with.rediation=as.factor(w$work.connected.with.rediation)
Socio.Econmic=as.factor(w$Socio.Econmic)
Kind.of.Vegetable=as.factor(w$Kind.of.Vegetable)
Kind.of.Meat=as.factor(w$Kind.of.Meat)
Sport=as.factor(w$Sport)
Smoke=as.factor(w$Smoke)
Drinking.Alcohol=as.factor(w$Drinking.Alcohol)
Other.diseases=as.factor(w$Other.diseases)
Inherited.diseases=as.factor(w$Inherited.diseases)
Marital.State=as.factor(w$Marital.State)
Family.history=as.factor(w$Family.history)
Age.at.menarche=as.factor(w$Age.at.menarche)
Age.at.first.Pregnancy=as.factor(w$Age.at.first.Pregnancy)
Breastfeeding=as.factor(w$Breastfeeding)
Length.of.Breastfeeding=as.factor(w$Length.of.Breastfeeding)
Number.of.chlidren=as.factor(w$Number.of.chlidren)
Spontaneous.abortions=as.factor(w$Spontaneous.abortions)
Age.at.last.pregnancy=as.factor(w$Age.at.last.pregnancy)
Age.at.menpause=as.factor(w$Age.at.menpause)
Oral.contraceptive.use=as.factor(w$Oral.contraceptive.use)
Duration.of.oral.contraceptive.use=as.factor(w$Duration.of.oral.contraceptive.use)
Place.of.Accommodation=as.factor(w$Place.of.Accommodation)
Breast.Cancer=as.factor(w$Breast.Cancer)
```

```
#####################################################################
##########################Demog factors ##################################
sim=50
S=9 # tree size
set.err=matrix(0,nc=sim,nr=S+2)
set.size=matrix(0,nc=sim,nr=S+2)
set.diff=matrix(0,nc=sim,nr=S+1)
set.sized=matrix(0,nc=sim,nr=S+1)
set.cum=matrix(0,nc=sim,nr=S+1)
 for(N in
1:sim){dat.dem=data.frame(Gender,Age,Education.level,Employee,Socio.Econmic,
Marital.State, Breast.Cancer)
n=nrow(w)
K=3 # number of folds
K.size=n/K
set.seed(N)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
all.err=numeric(0)
all.err.t=numeric(0)
size=numeric(0)
size=numeric(0)
for(k in 1:K){
for(s in 2: S){
dat.k=dat.dem[fold!=k,]
tre=tree(dat.k$Breast.Cancer~.,data=dat.k,method="class")
prun.tre= prune.tree(tre, best = s)
######### training error#######
pred.train=predict(prun.tre,newdata=dat.k,type="class")
tab.t=table(dat.k$Breast.Cancer,pred.train)
err.t=1-(tab.t[1,1]+tab.t[2,2])/sum(tab.t)
all.err.t=rbind(all.err.t,err.t)
```

```
pred.tre=predict(prun.tre,newdata=dat.dem[fold==k,],type="class")
tab=table(dat.dem$Breast.Cancer[fold==k],pred.tre)
err= 1-(tab[1,1]-Rab[2,2])/sum(tab)
all.err=rbind(err,all.err)
size=rbind(summary(prun.tre)$size,size)  }  }
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)      ##test error
sz=as.numeric(names(table(size)))
set.err[ 1:length(sz),N]=m.err
set.size[ 1:length(sz),N]=sz
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],8)
szd=sz[-l]
set.difff 1:length(szd),N]=diff
set.sized[ 1:length(szd),N]=szd
cum.diff=diff
 for(i in 2:length(szd)){  cum.diff[i]=cum.diff[i-l]+diff[i]}
set.cum[ 1:length(szd),N]=cum.diff}
## plot diff between test errors
plot(c(set.sized[set.sized!=0]),c(set.diff[set.sized!=0]), type="p", axes=F, ylab="Error
rate difference", xlab="Tree size",cex.lab=1.5)
z=c("2,3„„"3,4","4,5",„5,6","6,7„„"7,8","8,9","9,10")
axis(2,cex.axis=l.4)
axis(l, 3:10,z,cex.axis=l.l)
box()
#points(c(set.diff[set.sized!=0]))
m.diff=tapply(set.diff[set.sized!=0],set.sized[set.sized!=0],mean)
sz.diff=as.numeric(names(table(set.sized[set.sized!=0])))
lines(sz.diff,m.diff)
points(sz.diff,m.diff,pch=18,col=2)
diff between test errors
plot(c(set.sized[set.sized!=0]),c(set.cum[set.sized!=0]), type="p", axes=F, ylab="Error
rate difference", xlab="Tree size",cex.lab=1.5)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9","9,10")
```

```
axis(2,cex.axis=1.4)
axis(1, 3:10,z,cex.axis=1.1 )
box()
#points(c(set.cum[set.sized!=0]))
cum.diff=tapply(set.cum[set.sized!=0],set.sized[set.sized!=0],mean)
sz.diff=as.numeric(names(table(set.sized[set.sized!=0])))
lines(sz.diff,cum.diff)
points(sz.diff,cum.diff,pch=18,col=2)
##########################################################################
############################## Control factors ############################
sim=50
S=13 # tree size
set.err.con=matrix(0,nc=sim,nr=S+2)
set.size.con=matrix(0,nc=sim,nr=S+2)
set.diff.con=matrix(0,nc=sim,nr=S+1)
set.sized.con=matrix(0,nc=sim,nr=S+2)
set.cum.con=matrix(0,nc=sim,nr=S+2)
    for(N in 1:sim){
dat.con=data.frame(Weight,Kind.of.Vegetable,Kind.of.Meat,Sport,Breastfeeding,
Length.of.Breastfeeding,Number.of.chlidren,Age.at.last.pregnancy,Oral.contraceptive.u
se
,Duration.of.oral.contraceptive.use,Breast.Cancer)
n=nrow(w)
K=3
#S=12
K.size=n/K
set.seed(N)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
all.err.t=numeric(0)
all.err=numeric(0)
size=numeric(0)
```

```
for(k in 1:K){
for(s in 2: S){
dat.k=dat.con[fold!=k,]
tre=tree(dat.k$Breast.Cancer~.,data=dat.k,method="class")
prun.tre= prune.tree(tre, best = s)
        #### training error #####
pred.train=predict(prun.tre,newdata=dat.k,type="class")
tab.t=table(dat.k$Breast.Cancer,pred.train)
err.t=1-(tab.t[1,1]+tab.t[2,2])/sum(tab.t)
all.err.t=rbind(all.err.t,err.t)
 ############################ test error ##########
pred.tre=predict(prun.tre,newdata=dat.con[fold==k,],type="class")
tab=table(dat.con$Breast.Cancer[fold==k],pred.tre)
err=1-(tab[1,1]+tab[2,2])/sum(tab)
all.err=rbind(all.err,err)
size=rbind(size,summary(prun.tre)$size)}   }
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)     ##test error
sz=as.numeric(names(table(size)))
set.err.con[1:length(sz),N]=m.err
set.size.con[1:length(sz),N]=sz
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],8)
szd=sz[-1]
set.diff.con[1:length(szd),N]=diff
set.sized.con[1:length(szd),N]=szd
cum.diff=diff
 for(i in 2:length(szd)){ cum.diff[i]=cum.diff[i-1]+diff[i]}
set.cum.con[1:length(szd),N]=cum.diff}
## plot diff between test errors
plot(c(set.sized.con[set.sized.con!=0]),c(set.cum.con[set.sized.con!=0]), type="p",
axes=F, ylab="Cumulative error rate difference", xlab="Tree size",cex.lab=1.5)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9","9,10","10,11","11,12","12,13","13,14")
axis(2,cex.axis=1.4)
axis(1, 3:14,z,cex.axis=.7 )
```

```
box()
#points(c(set.cum.con[set.sized.con!=0]))
cum.diff=tapply(set.cum.con[set.sized.con!=0],set.sized.con[set.sized.con!=0],mean)
sz.diff=as.numeric(names(table(set.sized.con[set.sized.con!=0])))
lines(sz.diff,cum.diff)
points(sz.diff,cum.diff,pch=18,col=2)
#### plot test error and size ####
plot(size,all.err,type="p")
points(size,m.err,pch=20,col=4)
lines(size,m.err,col=2)
#### plot test and training error with size ###
size=as.numeric(names(table(size)))
plot(size,m.err.t,col=3,type="l",ylab="error rate")
points(size,m.err.t)
lines(size, m.err,col=2)
points(size,m.err)
######## tree with smallest test error  #####
t.demo=tree(dat.con$Breast.Cancer~.,data=dat.con,method="class")
prun.t= prune.tree(t.demo, best =8)
plot(prun.t);text(prun.t)
## plot diff between test errors
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],5)
sz=size[-1]
plot(diff, type="l", axes=F, ylab="Error rate difference", xlab="Tree size",cex.lab=1.5)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9","9,10","10,11","11,12")
axis(1, 1:10,z,cex.axis=1.1 )
axis(2,cex.axis=1.4)
box()
points(diff)
t.snip=snip.tree(prun.t,nodes=c(7))
plot(t.snip);text(t.snip)
```

```
set.err.uncon=matrix(0,nc=sim,nr=S+2)
set.size.uncon=matrix(0,nc=sim,nr=S+2)
set.diff.uncon=matrix(0,nc=sim,nr=S+1)
set.sized.uncon=matrix(0,nc=sim,nr=S+2)
set.cum.uncon=matrix(0,nc=sim,nr=S+2)
      for(N in l:sim){
dat.uncon=data.frame(Height,work.connected.with.rediation,Age.at.menarche,Age.at.fir
st.Pregnancy,Age.at.menpause,Spontaneous,abortions,Breast.Cancer)
n=nrow(w)
K=3
#S=8
K.size=n/K
set.seed(N)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
all.err.t=numeric(0)
all.err=numeric(0)
size=numeric(0)
for(k in 1:K){
for(s in 2: S){
dat.k=dat.uncon[fold!=k,]
tre=tree(dat.k$Breast.Cancer~.,data=dat.k,method="class")
prun.tre= prune.tree(tre, best = s)
#### training error #####
pred.train=predict(prun.tre,newdata=dat.k,type="class")
tab.t=table(dat.k$Breast.Cancer,pred.train)
err.t=l-(tab.t[ 1,1]+tab.t[2,2])/sum(tab.t)
all.err.t=rbind(all.err.t,err.t)
```

```
        #### test error #####
pred.tre=predict(prun.tre,newdata=dat.uncon[fold==k,],type="class")
tab=table(dat.uncon$Breast.Cancer[fold==k],pred.tre)
err=1-(tab[1,1]+tab[2,2])/sum(tab)
all.err=rbind(all.err,err)
size=rbind(size,summary(prun.tre)$size)}  }
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)     ##test error
sz=as.numeric(names(table(size)))
set.err.uncon[1:length(sz),N]=m.err
set.size.uncon[1:length(sz),N]=sz
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],8)
szd=sz[-1]
set.diff.uncon[1:length(szd),N]=diff
set.sized.uncon[1:length(szd),N]=szd
cum.diff=diff
 for(i in 2:length(szd)){ cum.diff[i]=cum.diff[i-1]+diff[i]}
set.cum.uncon[1:length(szd),N]=cum.diff }
## plot diff between test errors
plot(c(set.sized.uncon[set.sized.uncon!=0]),c(set.cum.uncon[set.sized.uncon!=0]),
type="p", axes=F, ylab="Cumulative error rate difference", xlab="Tree
size",cex.lab=1.5)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9")
axis(2,cex.axis=1.4)
axis(1, 3:9,z,cex.axis=1 )
box()
#points(c(set.cum.uncon[set.sized.uncon!=0]))
cum.diff=tapply(set.cum.uncon[set.sized.uncon!=0],set.sized.uncon[set.sized.uncon!=0]
,men)
sz.diff=as.numeric(names(table(set.sized.uncon[set.sized.uncon!=0])))
lines(sz.diff,cum.diff)
points(sz.diff,cum.diff,pch=18,col=2)
```

```
set.err.h=matrix(0,nc=sim,nr=S+2)
set.size.h=matrix(0,nc=sim,nr=S+2)
set.diff.h=matrix(0,nc=sim,nr=S+l)
set.sized.h=matrix(0,nc=sim,nr=S+2)
set.cum.h=matrix(0,nc=sim,nr=S+2)
      for(N in l:sim){
dat.health=data.frame(Other.diseases,Family.history,Inherited.diseases,
Breast.Cancer)
n=nrow(w)
K=3
#S=4
K.size=n/K
set.seed(N)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+l
fold=as.factor(fold)
print(summary(fold))
all.err.t=numeric(0)
all.err=numeric(0)
size=numeric(0)
for(k in 1:K){
for(s in 2: S){
dat.k=dat.health[fold!=k,]
tre=tree(dat.k$Breast.Cancer~.,data=dat.k,method="class")
prun.tre= prune.tree(tre, best = s)
 #### training error #####
pred.train=predict(prun.tre,newdata=dat.k,type="class")
tab.t=table(dat.k$Breast.Cancer,pred.train)
err.t=l-(tab.t[ 1,1 ]+tab.t[2,2])/sum(tab.t)
all.err.t=rbind(all.err.t,err.t)
```

```
        #### test error #####
pred.tre=predict(prun.tre,newdata=dat.health[fold==k,],type="class")
tab=table(dat.health$Breast.Cancer[fold==k],pred.tre)
err=1-(tab[1,1]+tab[2,2])/sum(tab)
all.err=rbind(all.err,err)
size=rbind(size,summary(prun.tre)$size) } }
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)     ##test error
sz=as.numeric(names(table(size)))
set.err.h[1:length(sz),N]=m.err
set.size.h[1:length(sz),N]=sz
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],8)
szd=sz[-1]
set.diff.h[1:length(szd),N]=diff
set.sized.h[1:length(szd),N]=szd
cum.diff=diff
 for(i in 2:length(szd)){ cum.diff[i]=cum.diff[i-1]+diff[i]}
set.cum.h[1:length(szd),N]=cum.diff}
## plot diff between test errors
plot(c(set.sized.h[set.sized.h!=0]),c(set.cum.h[set.sized.h!=0]), type="p", axes=F,
ylab="Cumulative error rate difference", xlab="Tree size",cex.lab=1.5)
#points(c(set.cum.h[set.sized.h!=0]))
cum.diff=tapply(set.cum.h[set.sized.h!=0],set.sized.h[set.sized.h!=0],mean)
sz.diff=as.numeric(names(table(set.sized.h[set.sized.h!=0])))
lines(sz.diff,cum.diff)
points(sz.diff,cum.diff,pch=18,col=2)
z=c("2,3","3,4","", "4,6","6,7")
axis(2,cex.axis=1.4)
axis(1, 3:7,z,cex.axis=1 )
box()
```

```
sim=50
S=13 # tree size
set.err.dh=matrix(0,nc=sim,nr=S+2)
set.size.dh=matrix(0,nc=sim,nr=S+2)
set.diff.dh=matrix(0,nc=sim,nr=S+1)
set.sized.dh=matrix(0,ne=sim,nr=S+l)
set.eum.dh=matrix(0,nc=sim,nr=S+1)

    for(N in l:sim){
dem.con=data.frame(Socio.Econmic,Education.level,Age,Employee,Weight,Sport,
Length.of.Breastfeeding,Kind.of.Vegetable,Breastfeeding,Breast.Cancer)
n=nrow(w)
K=3
#S=13
K.size=n/K
set.seed(N)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
all.err.t=numerie(0)
all.err=numeric(0)
size=numeric(0)
for(k in 1:K){
for(s in 2: S){
dat.k=dem.con[fold!=k,]
tre=tree(dat.k$Breast.Cancer~.,data=dat.k,method="class")
prun.tre= prune.tree(tre, best = s)
        #### training error #####
pred.train=predict(prun.tre,newdata=dat.k,type="class")
tab.t=table(dat.k$Breast.Cancer,pred.train)
err.t=l-(tab.t[ 1,1]+tab.t[2,2])/sum(tab.t)
```

```
all.err.t=rbind(all.err.t,err.t)
#### test error #####
pred.tre=predict(prun.tre,newdata=dem.con[fold==k,],type="class")
tab=table(dem.con$Breast.Cancer[fold==k],pred.tre)
err=1-(tab[1,1]+tab[2,2])/sum(tab)
all.err=rbind(all.err,err)
size=rbind(size,summary(prun.tre)$size) }}
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)      ##test error
sz=as.numeric(names(table(size)))
set.err.dh[1:length(sz),N]=m.err
set.size.dh[1:length(sz),N]=sz
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],8)
szd=sz[-1]
set.diff.dh[1:length(szd),N]=diff
set.sized.dh[1:length(szd),N]=szd
cum.diff=diff
 for(i in 2:length(szd)){ cum.diff[i]=cum.diff[i-1]+diff[i]}
set.cum.dh[1:length(szd),N]=cum.diff }
## plot diff between test errors
plot(c(set.sized.dh[set.sized.dh!=0]),c(set.cum.dh[set.sized.dh!=0]), type="p", axes=F,
ylab="Cumulative error rate difference", xlab="Tree size",cex.lab=1.5)
#points(c(set.cum.h[set.sized.h!=0]))
cum.diff=tapply(set.cum.dh[set.sized.dh!=0],set.sized.dh[set.sized.dh!=0],mean)
sz.diff=as.numeric(names(table(set.sized.dh[set.sized.dh!=0])))
lines(sz.diff,cum.diff)
points(sz.diff,cum.diff,pch=18,col=2)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9","9,10","10,11","11,12","12,13")
axis(2,cex.axis=1.4)
axis(1, 3:13,z,cex.axis=.8 )
box()
```

```
sim=50
S=13 #tree size
set.err.ch=matrix(0,nc=sim,nr=S+2)
set.size.ch=matrix(0,nc=sim,nr=S+2)
set.diff.ch=matrix(0,nc=sim,nr=S+1)
set.sized.eh=matrix(0,ne=sim,nr=S+l)
set.eum.ch=matrix(0,nc=sim,nr=S+1)
    for(N in l:sim){
uncon.hlth= data.frame(Height,work.connected.with.rediation,Spontaneous.abortions
,Family.history,Other.diseases,Inherited.diseases,Breast.Cancer)
n=nrow(w)
K=3
#S=9
K.size=n/K
set.seed(N)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
all.err.t=numerie(0)
all.err=numeric(0)
size=numeric(0)
for(k in 1:K){
for(s in 2: S){
dat.k=uncon.hlth[fold!=k,]
tre=tree(dat.k$Breast.Cancer~.,data=dat.k,method="class")
prun.tre= prune.tree(tre, best = s)
#### training error #####
pred.train=predict(prun.tre,newdata=dat.k,type="class")
tab.t=table(dat.k$Breast.Cancer,pred.train)
err.t=l-(tab.t[ 1,1]+tab.t[2,2])/sum(tab.t)
```

```
all.err.t=rbind(all.err.t,err.t)
        #### test error #####
pred.tre=predict(prun.tre,newdata=uncon.hlth[fold==k,],type="class")
tab=table(uncon.hlth$Breast.Cancer[fold==k],pred.tre)
err=1-(tab[1,1]+tab[2,2])/sum(tab)
all.err=rbind(all.err,err)
size=rbind(size,summary(prun.tre)$size)}  }
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)     ##test error
sz=as.numeric(names(table(size)))
set.err.ch[1:length(sz),N]=m.err
set.size.ch[1:length(sz),N]=sz
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],8)
szd=sz[-1]
set.diff.ch[1:length(szd),N]=diff
set.sized.ch[1:length(szd),N]=szd
cum.diff=diff
 for(i in 2:length(szd)){ cum.diff[i]=cum.diff[i-1]+diff[i]}
set.cum.ch[1:length(szd),N]=cum.diff }
## plot diff between test errors
plot(c(set.sized.ch[set.sized.ch!=0]),c(set.cum.ch[set.sized.ch!=0]), type="p", axes=F,
ylab="Cumulative error rate difference", xlab="Tree size",cex.lab=1.5)
#points(c(set.cum.h[set.sized.h!=0]))
cum.diff=tapply(set.cum.ch[set.sized.ch!=0],set.sized.ch[set.sized.ch!=0],mean)
sz.diff=as.numeric(names(table(set.sized.ch[set.sized.ch!=0])))
lines(sz.diff,cum.diff)
points(sz.diff,cum.diff,pch=18,col=2)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9","9,10","10,11","11,12","12,13")
axis(2,cex.axis=1.4)
axis(1, 3:13,z,cex.axis=.8 )
box()
```

```
set.err.all=matrix(0,nc=sim,nr=S+2)
set.size.all=matrix(0,nc=sim,nr=S+2)
set.diff.all=matrix(0,nc=sim,nr=S+1)
set.sized.all=matrix(0,nc=sim,nr=S+l)
set.cum.all=matrix(0,nc=sim,nr=S+1)
    for(N in l:sim){
all=data.frame(Weight,Sport,Length.of.Breastfeeding,Kind.of.Vegetable,Breastfeeding,
Height,Family.history,Breast.Cancer)
n=nrow(w)
K=3
#S=12
K.size=n/K
set.seed(N)
K.unif=runif(n)
range=rank(K.unif)
fold=(range-1)%/%K.size+1
fold=as.factor(fold)
print(summary(fold))
all.err.t=numeric(0)
all.err=numeric(0)
size=numeric(0)
for(k in 1:K){
for(s in 2: S){
dat.k=all[fold!=k,]
tre=tree(dat.k$Breast.Cancer~.,data=dat.k,method="class")
prun.tre= prune.tree(tre, best = s)
#### training error #####
pred.train=predict(prun.tre,newdata=dat.k,type="class")
tab.t=table(dat.k$Breast.Cancer,pred.train)
err.t=l-(tab.t[ 1,1]+tab.t[2,2])/sum(tab.t)
```

```
all.err.t=rbind(all.err.t,err.t)
            #### test error #####
pred.tre=predict(prun.tre,newdata=all[fold==k,],type="class")
tab=table(all$Breast.Cancer[fold==k],pred.tre)
err=1-(tab[ 1,1 ]+tab[2,2])/sum(tab)
all.err=rbind(all.err,err)
size=rbind(size,summary(prun.tre)$size)}  }
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)      ##test error
print(m.err)
print(m.err.t)
m.err.t=tapply(all.err.t,size,mean)  ## training error
m.err=tapply(all.err,size,mean)      ##test error
sz=as.numeric(names(table(size)))
set.err.all[ 1:length(sz),N]=m.err
set.size.all[ 1:length(sz),N]=sz
diff=m.err-c(m.err[-1],0)
diff=round(diff[ -length(diff) ],8)
szd=sz[-1]
set.diff.all[ 1:length(szd),N]=diff
set.sized.all[ 1:length(szd),N]=szd
cum.diff=diff
 for(i in 2:length(szd)){  cum.diff[i]=cum.diff[i-1]+diff[i]}
set.cum.all[ 1:length(szd),N]=cum.diff}
#.# plot diff between test errors
plot(c(set.sized.all[set.sized.all!=0 & set.sized.all!=13]),c(set.cum.all[set.sized.all!=0 &
set.sized.all!=13]), type="p", axes=F, ylab="Error rate difference", xlab="Tree
size",cex.lab=1.5)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9","9,10","10,11","11,12")
axis(2,cex.axis=1.4)
axis(1, 3:12,z,cex.axis=.9)
box()
#points(c(set.cum.all[set.sized.all!=0 & set.sized.all!=13]))
cum.diff=tapply(set.cum.all[set.sized.all!=0 &
set.sized.all!=13],set.sized.all[set.sized.all!=0 & set.sized.all!=13],mean)
```

```
sz.diff=as.numeric(names(table(set.sized.all[set.sized.all!=0 & set.sized.all!=13])))
lines(sz.diff,cum.diff)
points(sz.diff,cum.diff,pch=18,col=2)
##### plot test error and size ####
plot(size,all.err,type="p")
points(size,m.err,pch=20,col=4)
lines(size,m.err,col=2)
#### plot test and training error with size ###
size=as.numeric(names(table(size)))
plot(size,m.err.t,col=3,type="l",ylab="error rate")
points(size,m.err.t)
lines(size, m.err,col=2)
points(size,m.err)
## plot diff between test errors
diff=m.err-c(m.err[-1],0)
diff=round(diff[-length(diff)],5)
sz=size[-l]
plot(diff, type="l", axes=F, ylab="Error rate difference", xlab="Tree size",cex.lab= 1.5)
z=c("2,3","3,4","4,5","5,6","6,7","7,8","8,9","9,10","10,11","11,12","12,13")
axis(l, l:ll,z,cex.axis=l.l )
axis(2,cex.axis=l .4)
box()
```