Risk Factors and Classification of Diabetes in South Africa

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Nina Grundlingh

December, 2019

Risk Factors and Classification of Diabetes in South Africa

by

Nina Grundlingh

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
Master of Science
in
Statistics

Thesis Supervisor: Prof Temesgen Zewotir

Thesis Co-supervisor: Ms Danielle Roberts

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

UNIVERSITY OF KWAZULU-NATAL

School of Mathematics, Statistics and Computer Science

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Nina Grundlingh, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
   (a) their words have been re-written but the general information attributed to them has been referenced, or
   (b) where their exact words have been used, then their writing has been placed in italics and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

Nina Grundlingh (Student) Date

Prof Temesgen Zewotir (Supervisor) Date

Ms Danielle Roberts (Co-supervisor) Date

Disclaimer

This document describes work undertaken as a Masters programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

The Degree from which this study emanated was funded by the South African Medical Research Council under the Biostatistics Capacity Development funded in partnership with Building Academic Partnerships For Economics Development (BAPED): The content of any Publications from any studies during this Degree are solely the responsibility of the authors and do not necessarily represent the official views of the South African Medical Research Council.

Note

The following article is under peer review from this thesis:

- **Grundlingh, N.**, Zewotir, T., Roberts, D. & Manda, S. Assessment of prevalence and risk factors of diabetes and pre-diabetes in South Africa

The following have been presented from this thesis:

- **Grundlingh, N.**, Zewotir, T., Roberts, D. & Manda, S. Modelling risk factors of diabetes and pre-diabetes in South Africa. SUSAN-SSACAB 2019 Conference, 8-11 September 2019, Cape Town, South Africa.
- **Grundlingh, N.**, Zewotir, T., Roberts, D. & Manda, S. Modelling diabetes in the South African population. College of Agriculture, Engineering and Science Postgraduate Research & Innovation Symposium 2019, 17 October 2019, University of KwaZulu-Natal, Westville, South Africa (the award for best MSc presentation was also received for this).
- **Grundlingh, N.**, Zewotir, T., Roberts, D. & Manda, S. Modelling diabetes in South Africa. The 61st conference of the South African Statistical Association, 27-29 November 2019, Nelson Mandela University, South Africa

Abstract

Diabetes prevalence has been seen to be on the increase in recent years, globally and in South Africa. The number of people with diabetes globally has risen from 108 million in 1980 to 442 million in 2014. It was estimated that, of the 1.8 million people between 20 and 79 years old with diabetes in South Africa in 2017, 84.8% were undiagnosed. Diabetes was the 2nd leading underlying cause of death in South Africa in 2016. Identifying risk factors for diabetes will assist in raising public awareness and assist public authorities to develop prevention programs. This study aimed to investigate the prevalence and risk factors associated with diabetes in the South African population aged 15 years and older, as well as explore various statistical methods of classifying a person's diabetic status.

This study made use of the South African Demographic Health Survey 2016 data which involved a two-stage sampling design. The study participants included 6442 individuals aged 15 years and older. Of the individuals sampled, 11%, 67% and 22% were found to be non-diabetic, pre-diabetic and diabetic, respectively. Classification methods, namely, a decision tree, random forest and Bayesian neural network, were used to assess classification of diabetic status based on the risk factors. Of the classification methods, the Bayesian neural network gave the highest accuracy (75.9%). These methods however, failed to account for the complex survey design and sampling weights. In addition, these methods are not able to provide the estimated effect that a risk factor has on the diabetic status.

Regression models were employed to identify the significant risk factors. Due to the ordinal nature of diabetic status, initially the proportional odds model was fit. However, the proportional odds assumption was found to be violated. A multinomial generalized linear mixed model was fitted to account for the complexity of the design. However, the model's residuals were found to be spatially autocorrelated. Accordingly, a spatial generalized additive mixed model, which accounts for the complexity of the survey structure as well as incorporates nonlinear spatial effects, was adopted. The highest accuracy from the regression models considered was obtained from this adjusted surface correlation model (accuracy = 70.8%). Individuals of the Black/African race were more likely to be diabetic (OR = 1.429; 95% CI: 1.032-1.978) than other races. Individuals taking high blood pressure medication were 1.444 times more likely to be diabetic than pre-diabetic (95% CI: 1.167-1.786) compared to those not taking high blood pressure medication.

Acknowledgements

I want to thank my supervisors, Prof Temesgen Zewotir and Ms Danielle Roberts. Prof Zewotir for his knowledge and insight into applying statistics to real-world problems. I have learnt a great deal from Prof Zewotir from my Honours year and now in writing this thesis. Ms Roberts for all the encouragement, guidance and help, not only in writing this thesis.

Thank you to the SAMRC (South African Medical Research Council) for financial and academic support. I feel honoured to be associated with the high level of research outputs from the SAMRC. Thank you to the DHS (Demographic and Health Surveys) Program for allowing me access to their data else, this thesis would not have been possible.

My special thanks extended to the staff of the Department of Statistics at the University of KwaZulu-Natal. Each staff member has played a role in getting me thus far in my academic career and have inspired me to keep on pursuing academics. Thanks to the School's Dean, Prof Delia North, who heads this incredibly successful team.

Lastly, I would like to thank my family and friends who continue to support and encourage me in my academic endeavours.

Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CI	Confidence Interval
DHS	Demographic and Health Survey
DU	Dwelling Unit
EA	Enumeration Area
FN	False Negative
FP	False Positive
GDP	Gross Domestic Product
HbA1c	Glycated Heamoglobin
LCHF	Low-carbohydrate High-fat
IDF	International Diabetes Federation
OR	Odds Ratio
PSU	Primary Sampling Unit
SADHS	South African Demographic and Health Survey
SAMRC	South African Medical Research Council
SE	Standard Error
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
TN	True Negative
TP	True Positive
WHO	World Health Organisation

Chapter 1

Introduction

Diabetes, officially known as diabetes mellitus, is a chronic disease in which either not enough insulin is produced by the pancreas or the body cannot effectively make use of the insulin it produces. Three main types of diabetes exists: type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM) and gestational diabetes mellitus.

T1DM is due to an autoimmune disorder in which the cells in the pancreas do not produce or produce very little insulin. T1DM generally occurs in younger people but can occur at any age. Insulin injections are required to treat those with T1DM. T2DM is due to a metabolic disorder known as insulin resistance. T2DM is a progressive condition where no symptoms are apparent in the early stages. Insulin is commonly used for treatment however, insulin resistance worsens with insulin and thus causes further progression of T2DM (Noakes & Sboros, 2017). Lifestyle changes such as diet and physical activity are required in halting and even reversing progression of T2DM. T2DM was previously known as *adult onset diabetes* however, children as young as ten years old have been diagnosed with the disease (Reinehr, 2013). Gestational diabetes is hyperglycaemia which occurs in some women during pregnancy and usually disappears after pregnancy. Mother and child are then both at an increased risk in developing T2DM (IDF, 2019).

According to the World Health Organisation (WHO, 2018) the number of people with diabetes globally has risen from 108 million in 1980 to 442 million in 2014. Diabetes has been described as a ″new priority″ with middle- and low-income countries having seen a more rapid increase in the prevalence of diabetes.

The International Diabetes Federation (IDF, 2017) estimated that 73% of deaths due to diabetes in Africa in 2017 were people under 60 years of age. This was the highest percentage of all regions in 2017. South Africa had the second highest number of diabetes cases in Africa in 2017. Furthermore, of the 1.8 million people between 20-79 years old with diabetes in South Africa in 2017, 84.8% were undiagnosed. In 2016, tuberculosis was ranked the number 1, diabetes the 2nd and HIV as the 5th leading underlying cause of death in South Africa. Furthermore, diabetes was found to be the number one leading underlying cause of death for females in South Africa (STATS SA, 2016). According to the Indigo Wellness Index, South Africa was named the ″unhealthiest country on earth″ in 2019 (Millington, 2019). Thus, serious intervention is needed in order to improve South Africans' health.

The South African health system has been exhausted by endemics such as HIV and tuberculosis in recent years. Strategies such as the 90-90-90 target set for HIV/AIDS by UNAIDS in 2013 has seen great improvements in getting people tested for HIV, on treatment if they are tested positive and ultimately, getting those that are HIV-positive to a state where they are virally suppressed. As of 2018, it was estimated that South Africa was at 90-68-78 (AVERT, 2018). An increased life expectancy of 61 years old in 2010 to 67 years old in 2015 is due to South Africa having the largest antiretroviral treatment programme in the world (Mahlakoana, 2018). As of 2015, all those with HIV were put on antiretroviral therapy immediately, regardless of their CD4 count, based on the new WHO recommendations (Meintjes et al., 2017). Much progress has been made in South Africa in helping those with HIV know that they are infected and getting them on treatment. However, there is a lack of attention to diseases such as diabetes.

If diabetes is left untreated, serious nerve and blood vessel damage can occur, resulting in the following physical repercussions: eye problems, kidney damage (nephropathy), nerve damage (neuropathy), heart problems, foot problems, skin problems, teeth and gum problems, infections, thyroid problems and sexual dysfunction.

In the KwaZulu-Natal province of South Africa, 2500 leg amputations due to diabetes were performed in 2018 (Pijoos, 2018). Type 2 diabetics in South Africa are often diagnosed long after the disease has developed, when complications are dire and in-hospital stays are required (Green, 2017). This can negatively affect the country's economy. In 2015 it was calculated that the economic cost due to diabetes in sub-Saharan Africa was $19.5 billion or 1.2% of the gross domestic product (GDP), where these countries generally spend 5.5% of their GDP in total on health. This cost includes treatment, hospital stays and productivity losses due to early death, leaving the workforce or less productivity at work from poor health due to diabetes (The Lancet, 2017).

Diabetes can be diagnosed by one of the following tests:

- Glycated haemoglobin (HbA1c) test: measures your average blood glucose level for the previous 2-3 months. No fasting required.
- Fasting plasma glucose test: measures your immediate blood glucose level. Fasting for at least 8hrs is required prior to be tested.
- Oral glucose tolerance test: determines your body's efficiency in metabolising intake of sugar/carbohydrate. Blood is drawn to measure your blood glucose level after fasting for at least 8hrs. A liquid containing glucose is then consumed. Further blood samples are taken at regular intervals of 30 or 60 minutes and a single test is done after 2hrs. This test can take up to 3hrs.

1.1 Literature review

Diabetes remains underdiagnosed in South Africa. According to Peer et al. (2012), 57.9% of urban-dwelling black South Africans in Cape Town that were found to have diabetes were undiagnosed compared to the 52.2% that were undiagnosed in a similar study done almost 20 years prior by Levitt et al. (1993). Both studies included a 75-g oral glucose tolerance test.

Motala et al. (2008) studied diabetes and other glycaemia disorders in the rural Zulu district of Umbombo in KwaZulu-Natal. Individuals were classified as having diabetes, impaired glucose tolerance or impaired fasting glucose based on the 1998 WHO criteria for disorders of glycemia which involves a 75-g oral glucose tolerance test. A binary logistic regression model was used with a backward elimination method based on likelihood ratios in order to determine which variables were significantly associated with each health outcome being considered. The following variables were found to be associated with diabetes: family history of diabetes, alcohol consumption, waist and hip circumference, systolic blood pressure, levels of serum total triglycerides and total cholesterol. A moderate prevalence of diabetes and a high prevalence of total disorders of glycemia was found. There was no significant difference in prevalence for diabetes in men and women. Furthermore, peak age group prevalence was 55- to 64-year old.

There is uncertainty in which anthropometric measure is more associated with diabetes risk. There is evidence that measures of central obesity are more strongly associated with diabetes risk compared to body mass index (BMI) (Huxley et al., 2010). Waist circumference was found to be a risk factor for diabetes in the study done by Motala et al. (2008). It is believed that waist circumference had never been considered as a risk factor for diabetes in Africans prior to this study. There is thus a need for more research to be done to confirm this finding.

Peer et al. (2012) looked at psychosocial factors in a survey multiple logistic regression model to assess the prevalence of diabetes among urban-dwelling black South Africans. A separate regression model was obtained for men and women. Generalized additive models were used to assess the linearity of the variables under-consideration. From the men's regression model, the significant risk factors included older age, higher BMI and increasing waist circumference. The women's regression model obtained the same significant risk factors as well as low sense of coherence scores (mixture of optimism and control over one's environment), family history of diabetes and living in built formal housing. Low physical activity (<150min/week) was not found to be significantly associated with diabetes. Overall, there was an increased prevalence of diabetes and impaired glucose tolerance compared to the study done by Levitt et al. (1993) on a similar community. Peer et al. (2012) found the peak age-group for diabetes to be 65-74 year olds and thus a 10 year age gap when compared to the study done by Motala et al. (2008) on a rural black South African community.

Basu et al. (2013) were interested in determining whether obesity or sugar is the main driver of diabetes on a population-level. It was hypothesized that if obesity is the main driver hence measures would need to be put in place to reduce calorie consumption and increase physical activity. However, if it is found that added sugar consumption is the main driver then public health policies to reduce sugar consumption need to be put in place. The latter was found to be true. Sugar was the only food category (among fibre, fruit, meat, cereal and oils) to have a significant association with diabetes prevalence. Econometric models were applied and it was found that an increase of 150kcal/person/day (a can of soda) in sugar availability was associated with a 1.1% increase in diabetes prevalence after controlling for other food types, total calories, overweight and obesity, period effects, and several socio-economic variables.

Different diets have an effect on either the progression or remission of diabetes. Noakes (2013) conducted a survey on 127 individuals self-reporting on their weight change and overall health after adopting a low-carbohydrate, high-fat (LCHF) diet. In this study, 16 subjects reported that they no longer required medications for one or more of their medical conditions, the most common being type 2 diabetes (n=14), followed by hypertension (n=8) and then hypercholesterolaemia (n=7). Furthermore, 9 subjects with either T1DM or T2DM reduced their medications. This study emphasizes the fact that diabetes is reversible given the correct diet. Many other studies are in agreement with a LCHF diet having a reversible effect on diabetes (Malhotra et al., 2015; Feinman et al., 2015).

HIV-infected patients with cumulative exposure to combination antiretroviral therapy medications were found to have a increased incidence of diabetes. Specifically, stavudine and zidovudine have been found to be significantly associated to diabetes. Furthermore, HIV-infected patients who were currently smoking had a reduced risk of diabetes. Fasting plasma glucose was measured. Patients were defined to have definite new-onset diabetes if their fasting plasma glucose >7.0mmol/l (126mg/dl) on two consecutive occasions or possible diagnosis if a physician had reported a date of diabetes onset and initiated antidiabetic therapy (De Wit et al., 2008).

Classification methods have previously been used in disease classification, some of which will be noted here. Austin et al. (2013) assessed classification and predictions of heart failure subtypes in which it was found that tree-based methods performed the best. Ramani & Sivagami (2011) looked at finding the best accuracy classifier for Parkinson Disease. Different classification techniques including binary logistic regression, partial least squares regression, decision tree, random forest and support vector machine were applied to the data. The confusion matrix of each classification method was compared and it was found that the random forest yielded the highest accuracy (100%). Other studies on disease classification can be found in Kumari & Godara (2011); Kumari & Chitra (2013); MacGregor et al. (1994).

1.2 Thesis objectives

This study aimed at exploring various statistical methods for analysing data from a complex survey design by utilising the diabetes test results of the 2016 South African Demographic and Health Survey (SADHS) data. The specific objectives of the study are:

- To assess the prevalence of diabetes in individuals aged 15 years and older in South Africa
- To determine significant risk factors of diabetes in this sampled population
- To explore various statistical methods of classifying a person's diabetic status

1.3 Thesis outline

This thesis is organized in six chapters. Chapter 1 introduces the topic and the problem of diabetes in South Africa in particular.

Chapter 2 outlines the SADHS 2016 data set and some explanatory analyses are presented. In this chapter, we get an insight into the prevalence of diabetes in the sampled population and see how it differs according to demographic, health-related and lifestyle factors.

Chapter 3 highlights three classification techniques, namely the decision tree, random forest and Bayesian neural network. These techniques are then applied to the SADHS 2016 data to assess classification of diabetic status in the given population.

Chapter 4 discusses regression analysis for survey data. Here, we outline the generalized linear model before accounting for the survey design with the survey logistic regression model. The difference between an ordinal and nominal response is discussed. These regression methods are then applied to the SADHS 2016 data.

In Chapter 5, spatial statistics are introduced, specifically the problem with spatial autocorrelation in the residuals. We discuss how to correct for spatial autocorrelation and note some generalized additive mixed model theory. We then make the necessary correction for spatial autocorrelation observed in the SADHS 2016 data in Chapter 4.

In Chapter 6, a summary of the results is discussed. Conclusions and limitations of the study as well as possible areas of further study are given.

Chapter 2

Materials and Methods

2.1 The data set

South Africa is a country on the southernmost tip of the African continent and is comprised of 9 provinces. The South African population is made up of individuals with a wide variety of cultures, languages, and religions. The SADHS 2016 was designed to provide national, regional, urban and non-urban key estimates for the country as a whole. The survey was carried out from 27 June 2016 to 4 November 2016.

Questionnaires were based on the standard Demographic and Health Survey (DHS) questionnaires developed by The DHS Program. Modifications were made to consider the population and health issues applicable to South Africa. Five questionnaires were used: the Household Questionnaire, the Woman's Questionnaire, the Man's Questionnaire, the Caregiver's Questionnaire, and the Biomarker Questionnaire. The Household Questionnaire collected basic demographic information for each person listed. The Woman's Questionnaire collected information from woman 15-49 years old. The Man's Questionnaire collected information from men 15-59 years old. Both the Woman's and Man's Questionnaires included a module on adult health in which only one individual aged 15 years or older in the household answered. The adult health module included information on smoking, alcohol consumption, dietary habits, health care seeking behaviours, and self-reported prevalence of a variety of non-communicable diseases. The Biomarker Questionnaire recorded data on biomarkers such as anthropometry, anaemia testing, blood pressure measurements, HbA1c testing and HIV testing. This data was collected by trained nurses. HbA1c and HIV testing were only conducted on individuals 15 years and older. Furthermore, for adults 15 years and older, currently used prescribed medications were recorded. For the purpose of this study, only the Household Questionnaire, Woman's Questionnaire, Man's Questionnaire and the Biomarker Questionnaire were considered.

2.2 Sampling procedure

The Statistics South Africa Master Sample Frame was used. This involves the Census 2011 enumeration areas (EAs). EAs were then used as primary sampling units (PSUs). For survey precision, power allocation was used to allocate PSUs. Each province was stratified into urban, farm and traditional areas. This yielded 26 sampling strata, as the Western Cape does not have traditional residential geotype PSUs. The survey followed a stratified two-stage sampling design. At the first stage, a probability proportional to the size of PSUs was used where PSUs that contained more dwelling units had a higher chance of being selected. A total of 750 PSUs were selected from 26 sampling strata. This comprised of 468 PSUs in urban areas, 244 PSUs in traditional areas and 58 PSUs in farm areas. In January 2016 to March 2016, a list of all dwelling units was drawn up and used as a sampling frame for the selection of the dwelling units in the second stage. Systematic sampling was used to select a fixed number of 20 dwelling units per PSU/cluster. All dwelling units were subjected to the Household Questionnaire, the Woman's Questionnaire and the Caregiver's Questionnaire. In addition, the even numbered dwelling units were asked the Man's Questionnaire, the adult health module and had their biomarkers collected. The final sample consisted of 11083 households.

2.3 Data collection

Trained fieldworkers and nurses visited and interviewed the selected households. The necessary questionnaires were presented to each household.

Individuals 15 years and older were eligible for the HbA1c test. Finger-prick blood specimens were collected on a filter paper card by nurses. Blood samples were dried overnight and transported to the Global Clinical and Viral Laboratory the next morning. A blood chemistry analyser measures the total haemoglobin concentration by a colorimetric method. The HbA1c concentration was measured by a turbidimetric immunoinhibition method. HbA1c concentration is expressed as a percentage of total haemoglobin. The HbA1c measure is simple and convenient as it does not require one to be fasting. Thus, the HbA1c test has replaced the blood glucose test and oral glucose tolerance test. All these tests are considered accurate and acceptable. Glucose molecules attach themselves, in proportion to actual blood glucose levels, to red blood cells. The red blood cells carry haemoglobin and have a lifespan of three months. It is this attached glucose that is measured (Fung, 2018). The procedure and confidentiality of the data was explained to the respondents. Furthermore, respondents consented to the fact that the test results would not be made available to them.

2.4 Variables of interest

Interest is in the dependent, ordinal variable, diabetic status, indicating whether an individual is non-diabetic, pre-diabetic or diabetic. An individual is classified as pre-diabetic if their blood sugar is high but not high enough to be classified as diabetic. Thus, only these three categories for diabetic status exist (Fung, 2018).

The independent variables considered in modelling diabetes in South Africa in this thesis include demographic, health-related and lifestyle factors. These variables are given in Figure 2.1
The outcome variable in this study is the diabetic status of persons aged 15 years and older, categorised into three: non-diabetic, pre-diabetic and diabetic. The most important determinants of diabetes from various literature reviews (Motala et al., 2008; Huxley et al., 2010; Peer et al., 2012; Basu et al., 2013) were included, as well as those variables that were expected to be determinants. The explanatory variables at individual and household levels included gender, race, age, wealth category, individual's highest level of education, BMI category, Rohrer's index, waist circumference, waist-to-height ratio, blood pressure category, haemoglobin level, use of medication and blood pressure medication in specific, use of cigarette smoke in the previous 24hrs to being interviewed, perception of health, frequency of eating processed foods, approach towards salt consumption, and consumption of fruit, vegetables, fruit juice and sugary drinks the previous day.

2.5 Exploratory data analysis

This section serves to assess the nature and characteristics of the data with which we are dealing. Since we are interested in the respondents' HbA1c measure, only those individuals that had the HbA1c test completed are considered in our analysis. There were 3636 households made up of 6442 individuals that fully completed both the HbA1c test and the adult health module.

It should be noted that only one individual per household answered the adult health module on behalf of the household. Thus, for the purpose of this study, it is assumed that their response holds true to other members of the household that agreed to the

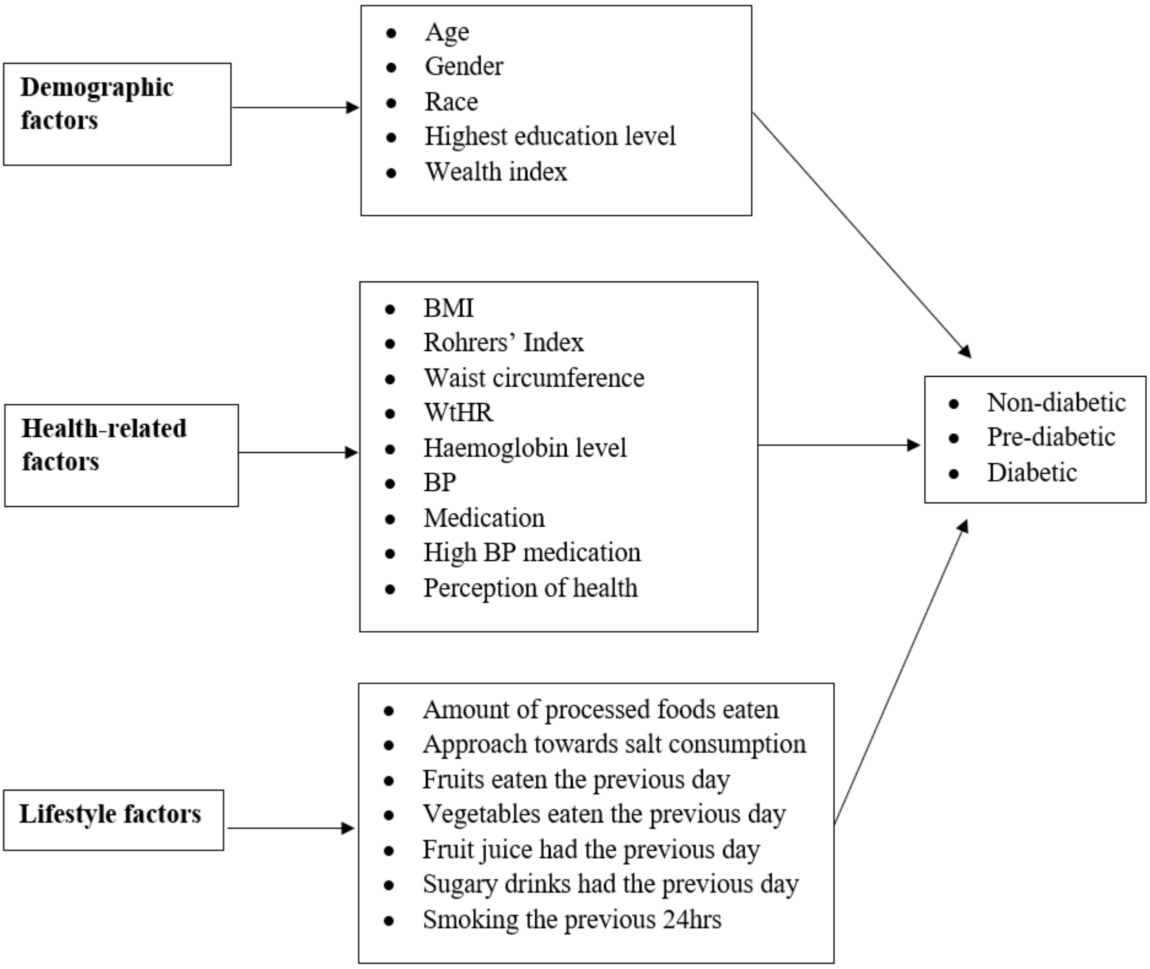

Figure 2.1: Conceptual framework of variables of interest

HbA1c testing. The wealth index was calculated by means of dividing the households into quintiles of poorest, poorer, middle, richer, richest, based on their wealth index Z-score. Approach towards salt consumption was considered positive if individuals have or believe they should reduce their salt intake and negative otherwise.

Table 2.1 shows the distribution of counts of diabetic status for different categorical variables of interest. Of the individuals that have primary school as their highest level of education, 69.2% are pre-diabetic. It can be seen that a high percentage of individuals taking high blood pressure medication are diabetic (41.2%) as well as those taking any medication in general (37.1%). Of those individuals that believe to have an excellent perception of health, 71.1% are pre-diabetic.

Table 2.1: Counts across diabetic status for different categorical variables

Variable of interest	Non-diabetic	Pre-diabetic	Diabetic
Gender			
Female	416 (10.4%)	2597 (64.9%)	989 (24.7%)
Male	326 (13.4%)	1695 (69.5%)	419 (17.2%)
Race			
Black/African	632 (11.1%)	3840 (67.3%)	1235 (21.6%)
Other	110 (15%)	452 (61.5%)	173 (23.5%)
Highest education level			
Primary	528 (12.9%)	2824 (69.2%)	730 (17.9%)
Secondary	155 (8.4%)	1126 (60.8%)	572 (30.9%)
Other	59 (11.6%)	342 (67.5%)	106 (20.9%)
Wealth category			
Poor	146 (12.2%)	803 (66.9%)	252 (20.1%)
Middle	138 (10.7%)	897 (69.4%)	258 (20%)
Rich	458 (11.6%)	2592 (65.7%)	898 (22.8%)
BMI category			
Underweight	79 (20.5%)	270 (69.9%)	37 (10.0%)
Normal	399 (15.1%)	1909 (72.0%)	342 (12.9%)
Overweight to severely obese	264 (7.8%)	2113 (62.0%)	1029 (30.2%)
Blood pressure category			
Normal	538 (13.2%)	2813 (69.2%)	714 (17.6%)
Abnormal	204 (8.6%)	1479 (62.2%)	694 (29.2%)
Taking high blood pressure medication			
No	672 (12.9%)	3634 (69.8%)	897 (17.2%)
Yes	70 (5.6%)	658 (53.1%)	511 (41.2%)
Taking medication			
No	661 (12.9%)	3549 (69.2%)	921 (17.9%)
Yes	81 (6.2%)	743 (56.7%)	487 (37.1%)
Health perception			
Poor	92 (9.9%)	586 (62.9%)	253 (27.2%)

Table 2.1 – *Continued from the previous page*

Variable of interest	**Non-diabetic**	**Pre-diabetic**	**Diabetic**
Average	259 (10.9%)	1561 (65.9%)	548 (23.1%)
Good	314 (12.7%)	1660 (67.3%)	493 (20.0%)
Excellent	77 (11.4%)	485 (71.1%)	114 (16.9%)
Ate fruit yesterday			
Yes	318 (10.7%)	1962 (66.3%)	679 (22.9%)
No	424 (12.3%)	2300 (66.6%)	729 (21.1%)
Ate vegetables yesterday			
Yes	418 (11.0%)	2485 (65.6%)	885 (23.4%)
No	324 (12.2%)	1807 (68.1%)	523 (19.7%)
Approach towards salt consumption			
Positive	502 (11.1%)	2992 (66.1%)	1034 (22.8%)
Negative	240 (12.5%)	1300 (67.9%)	374 (19.5%)
Had a sugary drink yesterday			
Yes	247 (11.8%)	1408 (67.2%)	441 (21.0%)
No	495 (11.4%)	2884 (66.4%)	967 (22.3%)
Had fruit juice yesterday			
Yes	108 (13.3%)	515 (63.3%)	190 (23.4%)
No	634 (11.3%)	3777 (67.1%)	1218 (21.6%)
Smoked cigarettes the previous 24hrs			
Yes	151 (15.4%)	686 (70.1%)	141 (14.4%)
No	591 (10.8%)	3606 (66.0%)	1267 (23.2%)

Figure 2.2 displays the percentage of individuals that tested non-diabetic, pre-diabetic and diabetic. The figure shows that 22% of the individuals are diabetic and an alarmingly high 66% are pre-diabetic. This 66%, if left untreated, could develop diabetes and face the consequences thereof.

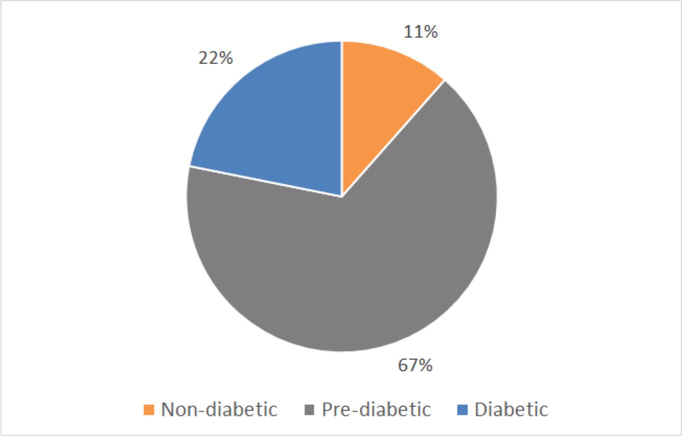

Figure 2.2: Observed percentage of diabetic status

Figure 2.3 shows the diabetic status for males and females across the different age groups. Considering that of the females, there appears to be a decreasing trend of non-diabetics and pre-diabetics. There is an increasing trend of diabetics. This suggests that females have an increasing HbA1c measure with age and hence, diabetes occurs more commonly later in life. The 75-100 year age group has the highest count of diabetics and lowest count of non-diabetics (ignoring the 65-69 year old age group in which no individuals were counted). The 25-29 year old age group has the highest count of pre-diabetics.

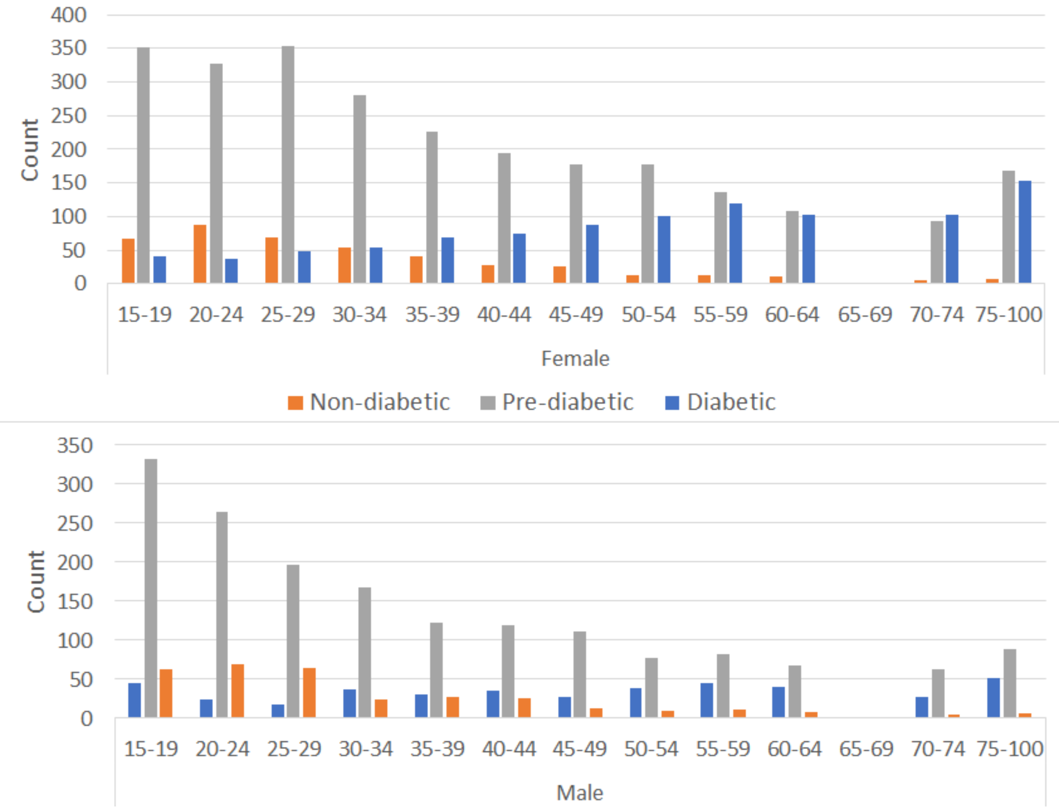

Figure 2.3: Diabetic status across different age groups for females and males

Similar to that of the females, there is a decreasing trend of non-diabetics among men (Figure 2.3). However, an increasing trend of diabetics isn't as apparent as is seen among the females. The 75-100 year age group, again, has the highest count of diabetics and the 15-19 year old age group has the highest count of pre-diabetics. The large number of youngs males and females (15-29 year old) with pre-diabetes is indicative of a high prevalence of diabetes in the future if these individuals are left untreated.

As seen in Figure 2.4 which displays the distribution of the highest education level of individuals for each diabetic status, 71% of non-diabetic individuals have secondary education as their highest level of education, 15% have primary school, 8% have higher education and 6% have no education. Considering highest education of pre-diabetics, 66% have secondary education as their highest level of education, 18% have primary school, 8% have higher education and 8% have no education. An increasing trend can be seen in no education and primary school education as the diabetic status of the individuals worsens. Also, there is a decreasing trend in secondary school as the highest level of education as the diabetic status of individuals worsens. It is important to note with this data that the largest count of individuals is coming from the 15-19 age group (as seen in Figure 2.3) hence, it is not likely that many of these individuals would have reached higher education, or even completion of secondary education, at the time that this survey was conducted. In all three diabetic statuses (Figure 2.4), the same percentage of individuals with higher education is observed (8%).

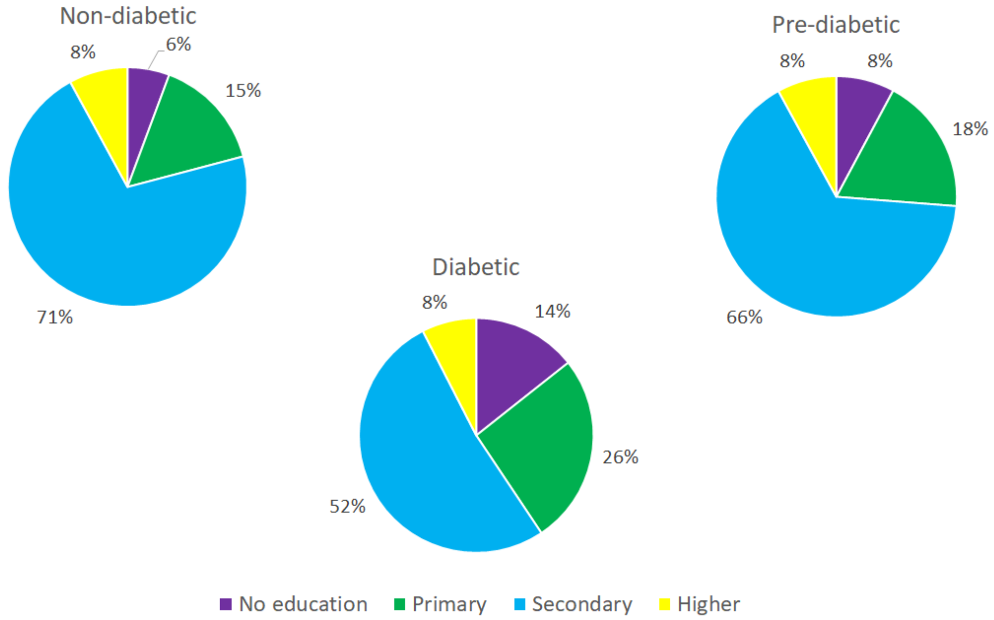

Figure 2.4: Spread of highest level of education according to diabetic status

The WHO has a guideline for BMI categories in order to classify an individual's weight status (see Table 2.2).

Table 2.2: WHO guidelines to BMI

Status	Value (kg/m^2)
Underweight	< 18.5
Normal	$18.5 - 24.9$
Overweight	25.0-29.9
Obese	≥ 30.0
Severely obese	≥ 35.0

Figure 2.5 displays the mean BMI of males and females across diabetic status. The red line marks the minimum BMI for which an individual would be categorised as overweight according to the WHO standards given in Table 2.2. The mean BMI of males and females follows a similar trend in that the diabetics have the highest mean BMI followed by pre-diabetics and then non-diabetics. Overall, females have a higher mean BMI than the males with females from all diabetic categories having a mean BMI that would be classified as overweight by the WHO. Among the males, only the diabetics have a higher probability of being classified as overweight or worse.

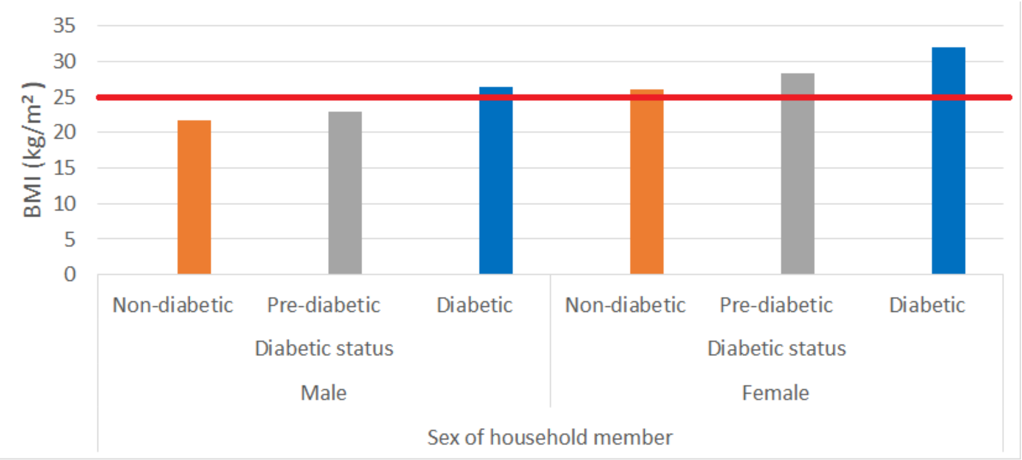

Figure 2.5: Mean BMI of males and females across diabetic status

We will not only consider BMI, as there exists more anthropometric measures which give us an indication of an individual's body composition, such as Rohrer's Index, waist circumference and waist-to-height ratio. The average of these measures across the different diabetic statuses for males and females is shown in Figures 2.6 to 2.8. Similarly, as was seen with mean BMI in Figure 2.5, the mean Rohrer's Index for diabetics is the highest followed by pre-diabetics and then non-diabetics having the lowest mean Rohrer's Index (Figure 2.6 on the next page). The males, on average, have a lower mean Rohrer's Index across the different diabetic statuses when compared to the females.

Note that individuals across all the diabetic statuses have a mean Rohrer's Index greater than 12kg/m^2 (noted by the red line) which indicates an increase in risk of metabolic complications. Again with Figure 2.7, the mean waist circumference for diabetics is the greatest followed by pre-diabetics and then non-diabetics. Also, males appear to have a lower mean waist circumference than females.

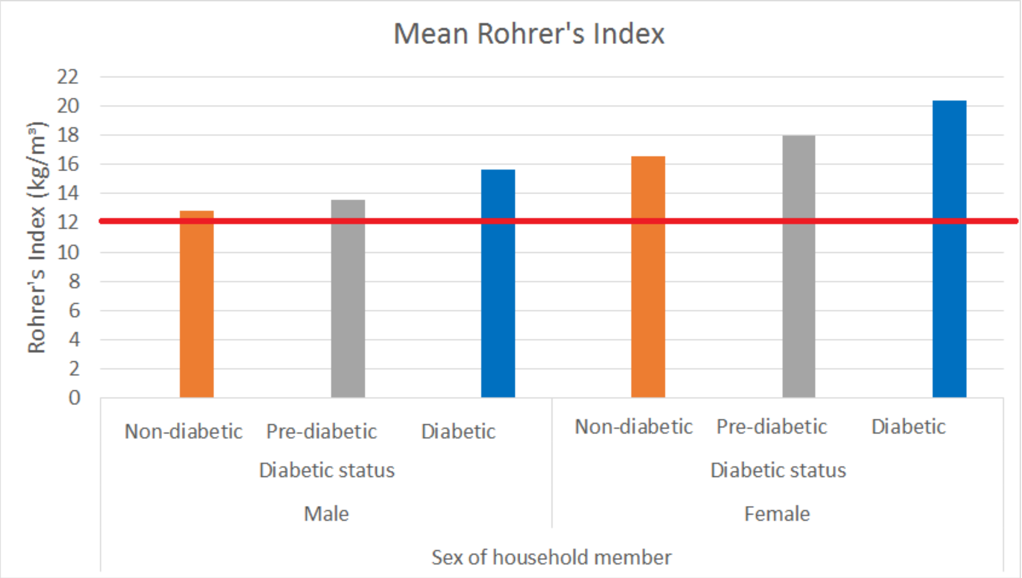

Figure 2.6: Mean Rohrer's index for males and females across diabetic status

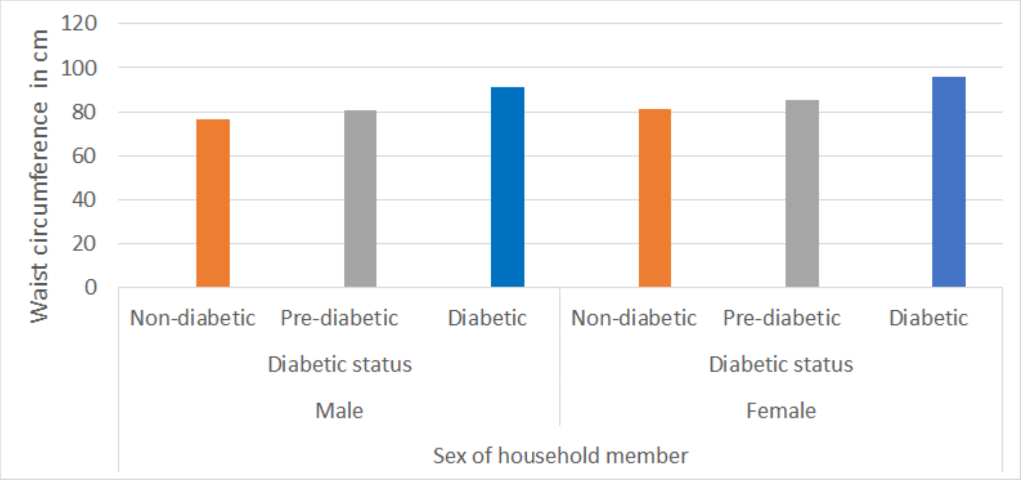

Figure 2.7: Mean waist circumference for males and females across diabetic status

In Figure 2.8, the same trend continues as was seen in the above Figures 2.5, 2.6 and 2.7. Thus, all anthropometric measurements show diabetics having the greatest mean measure followed by pre-diabetics and then non-diabetics. As well as males having a lower mean measurement compared to females. The red line in Figure 2.8 indicates the lower limit for those being at risk for metabolic complications (0.5) (Ashwell & Gibson, 2016).

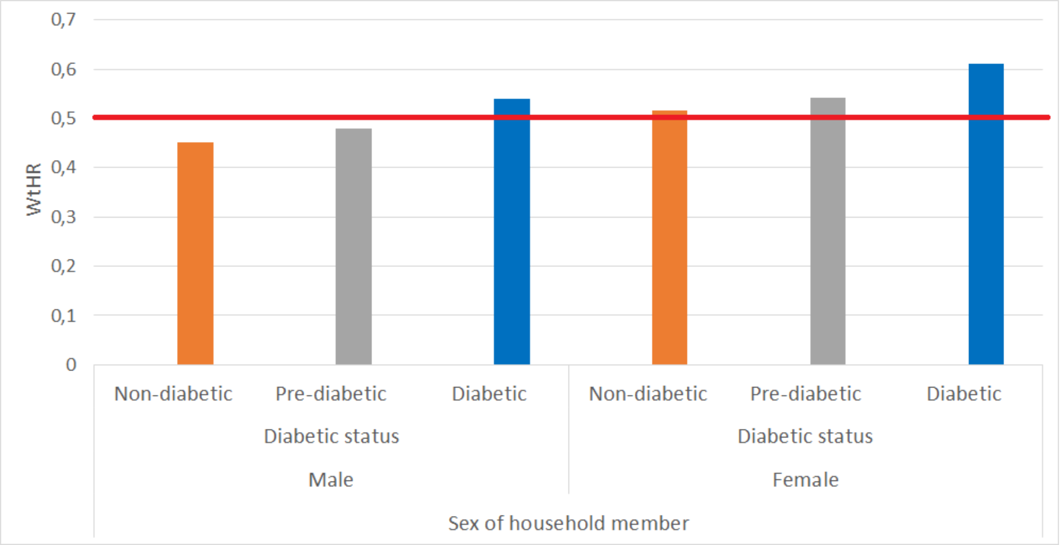

Figure 2.8: Mean waist-to-height ratio for males and females across diabetic status

Figure 2.9 displays the mean haemoglobin level, adjusted for altitude and smoking, across the different diabetic statuses. There is a slight decrease in mean haemoglobin level from non-diabetics to diabetics however, this is a very small difference and there is not much difference in haemoglobin levels as a marker.

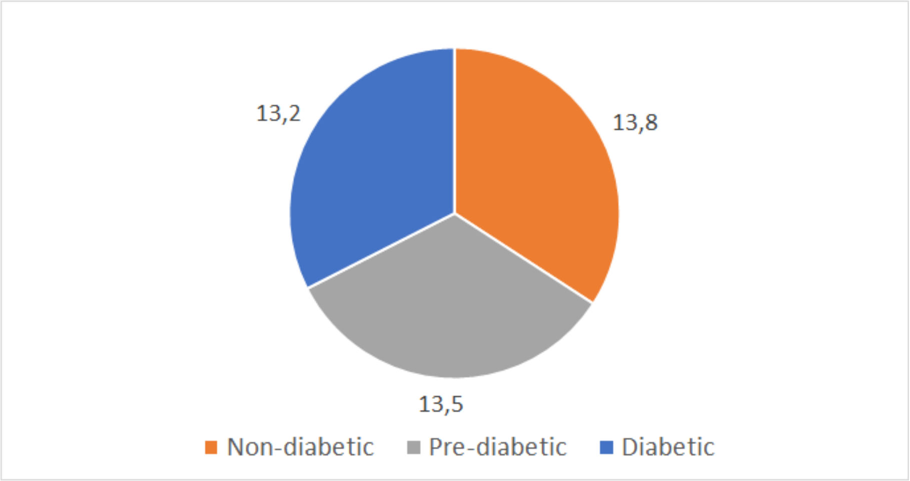

Figure 2.9: Mean haemoglobin levels adjusted for altitude and smoking in g/dl

Figure 2.10 on the next page gives the percentage of those with normal and abnormal blood pressure for each diabetic status. An individual with any elevated blood pressure is considered as having abnormal blood pressure. This figure displays an increasing trend in the percentage of individuals with abnormal blood pressure as their diabetic status worsens, with 27% of non-diabetics, 34% of pre-diabetics and 49% of diabetics with abnormal blood pressure. Thus, there appears to be some association between diabetes and blood pressure.

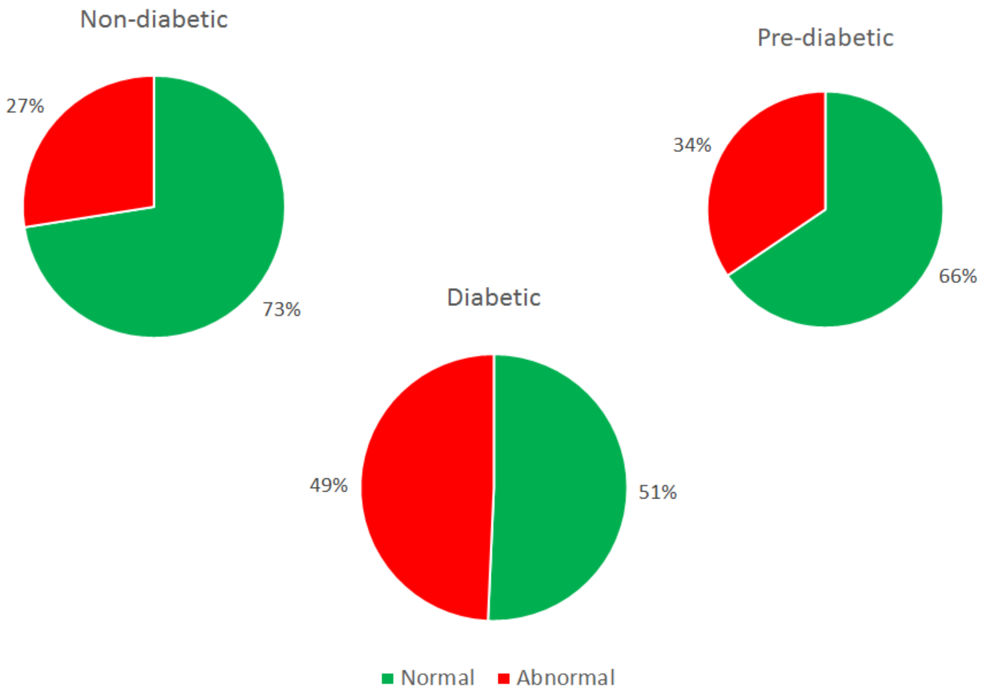

Figure 2.10: Percentage of non-diabetics, pre-diabetics and diabetics with normal and abnormal blood pressure

From Figure 2.11, a very small percentage (20% and less) in each category smoked in the previous 24 hours to being asked this question. Non-diabetics had the highest percentage of those smoking, with a decreasing trend as the diabetic status worsens.

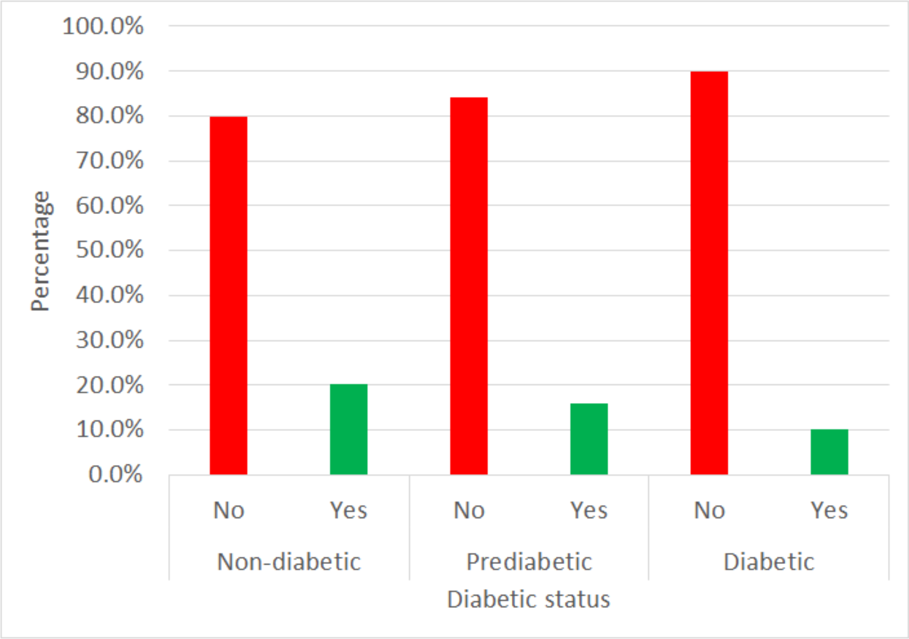

Figure 2.11: Percentage in each diabetic status of those that smoked in the previous 24hrs

Figures 2.12 and 2.13 show the percentage of individuals who drank a sugar-sweetened drink and/or fruit juice on the previous day to being interviewed across the different diabetic statuses. From Figure 2.12, there is almost an even percentage (around 32%) throughout the different diabetic statuses of those that drank a sugar-sweetened drink the previous day to being interviewed. No diabetic status had individuals who appeared to consume more than the others. Similarly, the percentage of those that drank fruit juice the previous day is almost even (around 12%) across the different diabetic statuses (Figure 2.13).

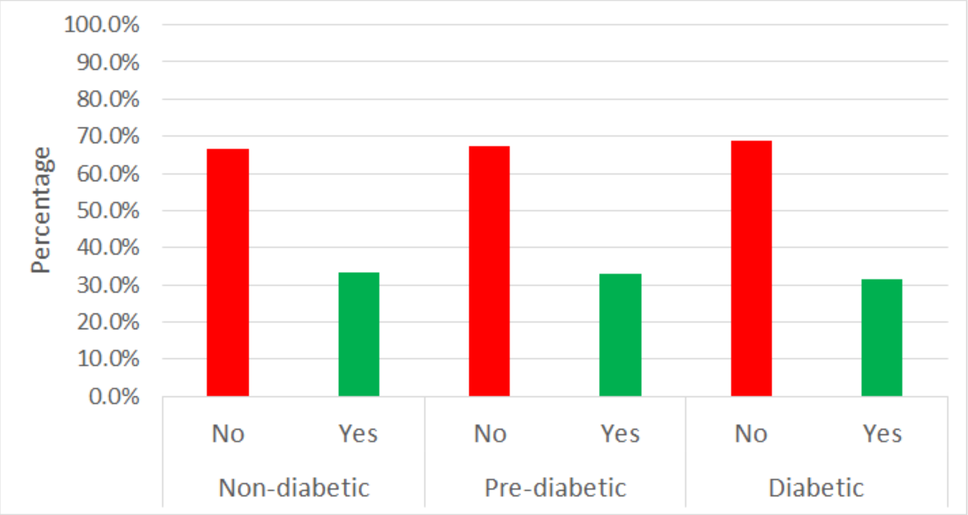

Figure 2.12: Percentage of those that drank sugar-sweetened drinks the previous day across diabetic status

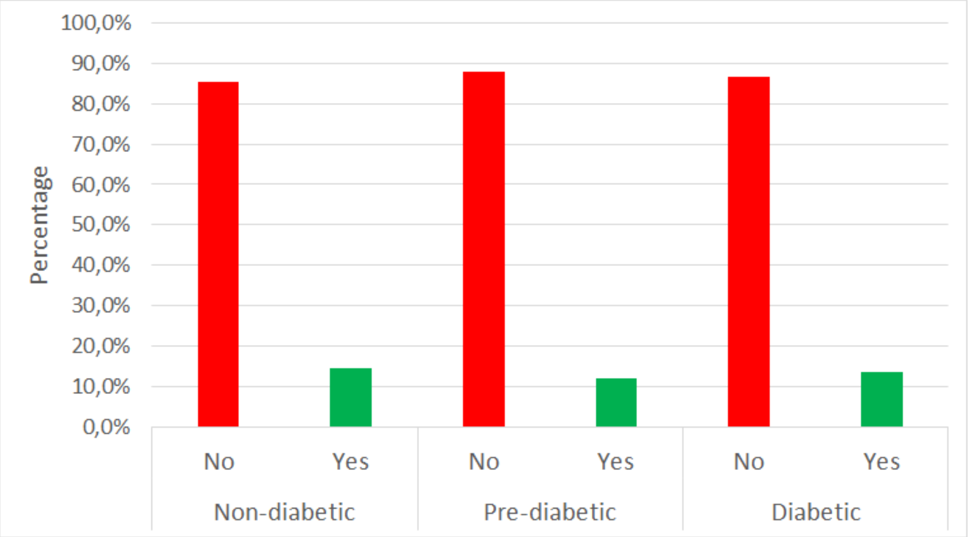

Figure 2.13: Percentage of those that drank fruit juice drinks the previous day across diabetic status

Figure 2.14, which displays an individual's perception of health according to their diabetic status, shows that a higher percentage (42%) of non-diabetics have a good perception of health compared to pre-diabetics and diabetics. Most pre-diabetics have a good perception of health (39%) but not as many as non-diabetics. Diabetics have a higher percentage of individuals with a poor (18%) or average (39%) perception of health compared to non-diabetics and pre-diabetics.

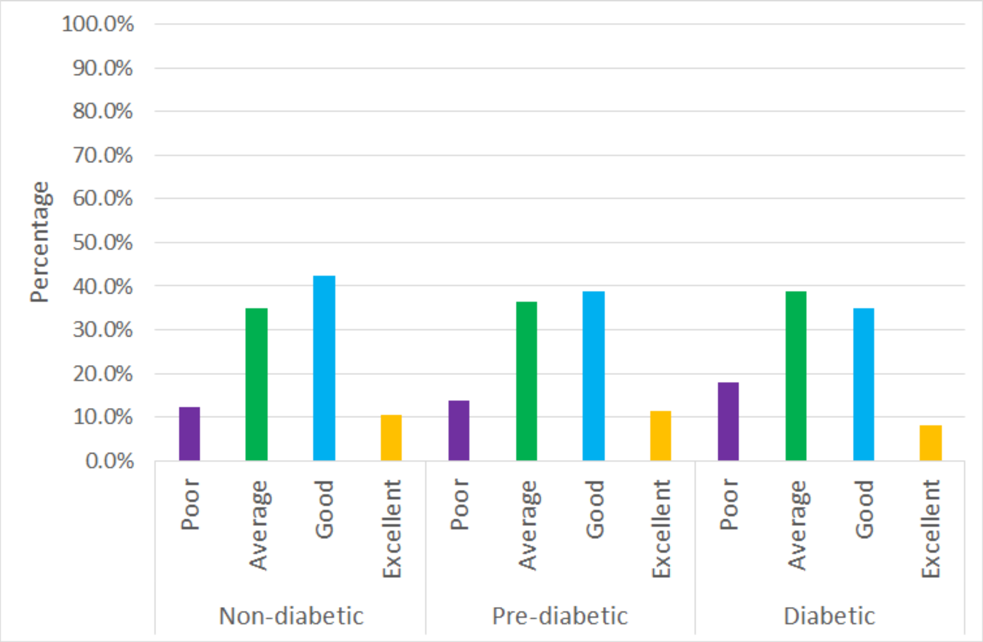

Figure 2.14: Perception of health across the different diabetic status

Rather than considering the frequency in which packed chips, fast food, fried food and processed meat were each eaten, it was of interest to only know if they were eaten. Each of the four were coded as a binary variable, '0' for never eaten and '1' for eaten. Thus, responses 'every day', 'at least once a week' and 'occasionally' were coded as '1'. A new variable, processed food, was then created by summing across the four foods and thus giving a total, out of 4, of the foods that were eaten. This gives us an indication of how much variety of processed foods a household had eaten and was thus treated as a continuous variable. Figure 2.15 gives a summary of the variety of processed foods consumed by individuals with each diabetic status. There is not much difference seen across diabetic status with each status consuming on average 3 varieties of processed foods.

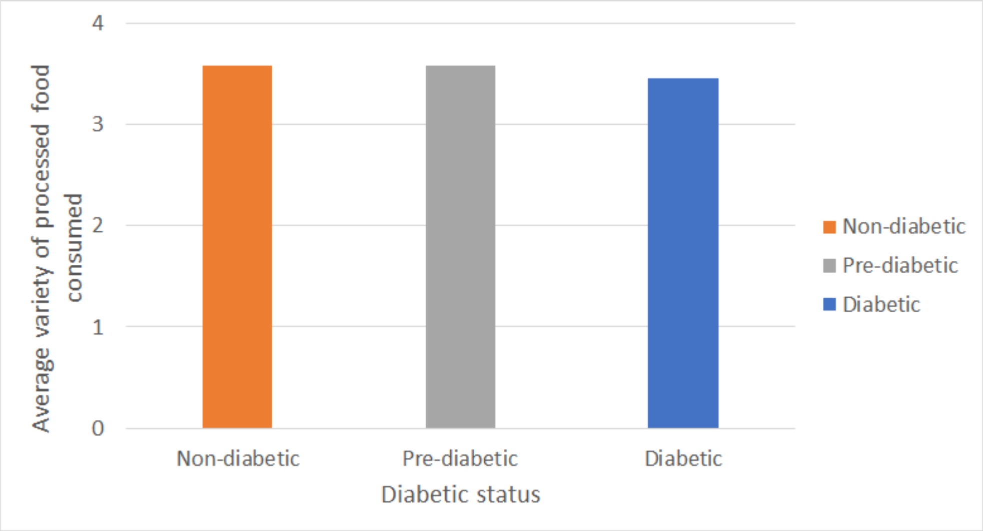

Figure 2.15: Average variety of processed food consumed across diabetic status

The parallel plot displayed in Figure 2.16 on the next page, gives a visualisation of the continuous variables of interest in relation to diabetic status. Non-diabetics appear to be less apparent in older age groups. Diabetics and pre-diabetics appear to have a higher Rohrer's index and waist circumference whereas, non-diabetics appear to be lower in these anthropometric measures, as also seen in Figures 2.6 to 2.8 earlier. On the whole, individuals have a low waist-to-height, ratio with diabetics and pre-diabetics appearing in the upper range. A range of haemoglobin levels is seen among diabetics, pre-diabetics and non-diabetics.

From this exploratory analysis, we get a sense of which variables play a role in classifying one's diabetic status. What follows in the next chapter is statistical classification methods to examine if the variables are able to classify diabetic status.

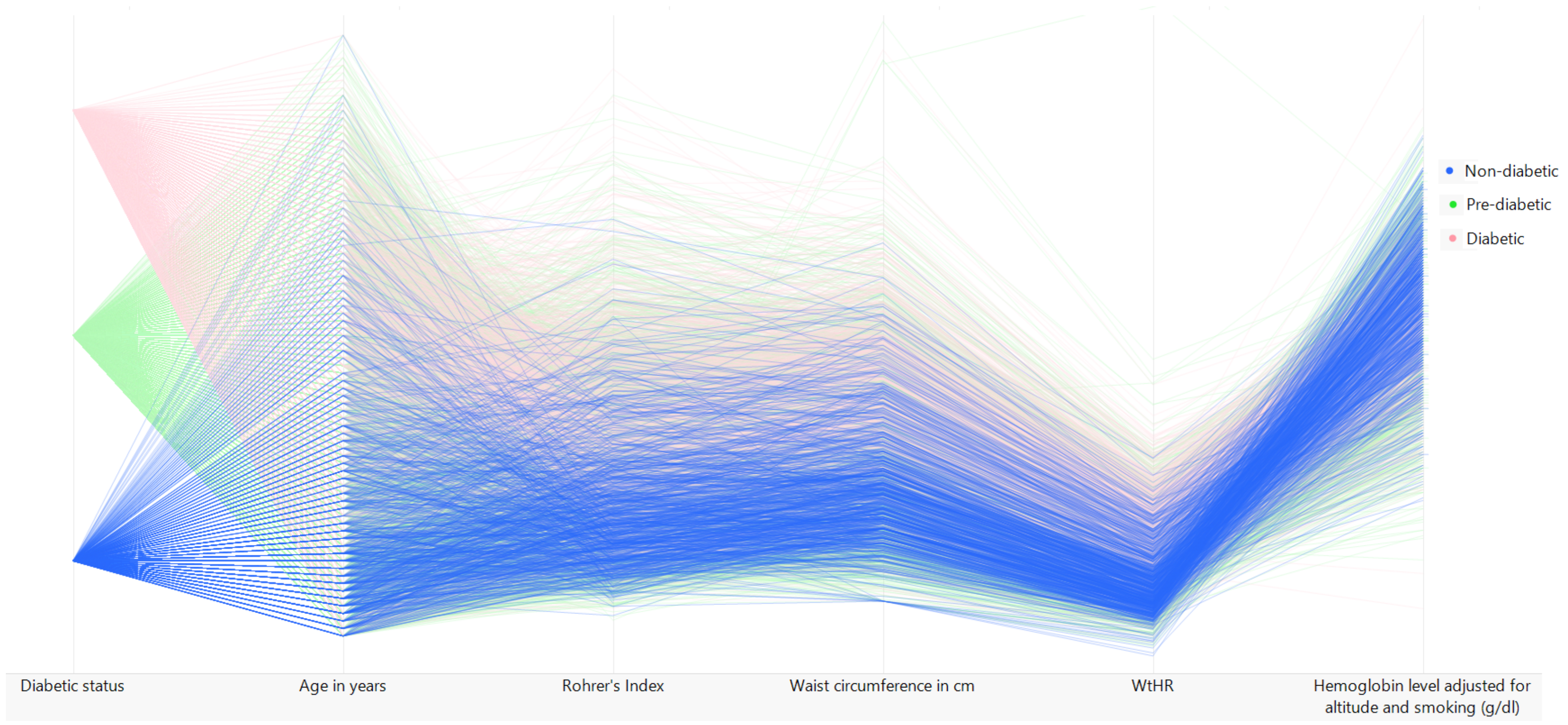

Figure 2.16: Parallel plot to show the relationship between diabetic status and some continuous variables

Chapter 3

Classification Methods

Classification methods are used to learn a sample from past experience where the measurement data consists of N cases observed in the past along with their classification. The purpose of classification analysis is to produce an accurate classifier or to determine the predictive structure of a problem. By determining the predictive structure, we are thus interested in the variables or interactions that determines when an object is in a specific class and not another. In this chapter we consider three different classification methods and assess how they perform on determining diabetic status from the risk factors given in Chapter 2.

3.1 Decision trees

Decision trees are one of the best supervised learning algorithms meaning that it deals with a pre-defined target variable. A decision tree has a flowchart-like structure which supports modelling decisions. A decision tree starts at the top with a root node that branches out to internal and leaf nodes. Each internal node represents a point of decision, a branch represents the outcome of a decision, each leaf node represents the possible classification or decision taken. When splitting a node, every feature is considered but the one that produces the most separation between observations is selected Breiman et al. (1984).

A decision tree begins with all observations which are then split into two groups according to the best value of any independent variable. We then have two child nodes. At each node, we want a feature that will split the observations so that each group is as different from each other as possible while the members in each group are as similar to each other as possible. The split is determined by finding the best independent variable to split on and a best cutpoint. The splitting criterion can be governed by maximising the decrease in node impurity or based on a statistical test.

The impurity of a parent node, $i(\tau)$, can be defined as a nonnegative number that is equal to zero for a pure node, a pure node being one in which all observations have the same value in the response variable. It is desired to produce the highest reduction in impurity

$$\delta i(s, \tau) = i(\tau) - \sum_{b=1}^{B} p(\tau_b|\tau) i(\tau_b) \tag{3.1}$$

where τ_b denotes the b^{th} child node, $p(\tau_b|\tau)$ is the proportion of observations in τ that are assigned to τ_b, and B is the number of branches after splitting τ.

The different impurity reduction criteria are:

- Entropy criterion
  The entropy impurity of node τ is defined as

  $$i(\tau) = -\sum_{j=1}^{J} p_j \log_2 p_j \tag{3.2}$$

  where p_j is the proportion of observations that have the j^{th} response value.

- Gini index criterion
  Here $i(\tau)$ is defined as the Gini index that corresponds to the average square error (ASE) of a class response and is given by

  $$i(\tau) = 1 - \sum_{j=1}^{J} p_j. \tag{3.3}$$

- Residual sum of squares criterion
  This impurity reduction criterion is used in regression trees whereas the entropy and Gini index criterion are used for classification trees. The impurity of node τ is defined as the residual sum of squares

  $$i(\tau) = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} (Y_i - \bar{Y})^2 \tag{3.4}$$

  where $N(\tau)$ is the number of observations in τ, Y_i is the response value of observation i, and $\bar{Y}$ is the average response of the observations in τ.

The criteria based on statistical tests include the chi-square criterion for categorical response variables, F-test for continuous response variables or the CHAID criterion which can be used for both categorical and continuous response variables. Furthermore, for categorical response variables there is also the FastCHAID criterion.

Three different algorithms are implemented to generate candidate splits. First, the exhaustive method is implemented. If the number of computations exceeds the threshold specified, the greedy algorithm is implemented. Likewise, if the number of computations again exceeds the threshold, the fast-sort method is implemented. Note that a variable can be used more than once in a branch as long as the split is on a different value of that variable.

Now, once the tree is fully grown there is potential of overfitting the data due to its large size. The tree could be too specific to the data it has been established from and thus not generalize well to new data. To counter this problem a smaller subtree must be established that is low in error rate. However, it must not be so small that it fails to capture important structural information. This is achieved by pruning the tree to get the optimal subtree.

Here we describe the cost-complexity pruning method. This algorithm is built around the trade-off of the complexity of a tree and the error rate to prevent overfitting. For a tree T the cost-complexity $R_\alpha(T)$ is given by

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{3.5}$$

where the error rate is represented by $R(T)$, the number of leaves on tree T is $|\tilde{T}|$ and α is the complexity parameter which gives the cost of each leaf. For a categorical response, misclassification rate is used for the error rate $R(T)$. In pruning, the aim is to minimise this function. If $\alpha = 0$ then this is the full tree T_0 with all nodes intact. As α increases the corresponding subtree gets smaller until we are left with a tree of size 1. In general, we want the subtree $T(\alpha)$ that minimises $R_\alpha(T)$

$$R_\alpha(T(\alpha)) = \min_{T \leq T_{max}} R_\alpha(T). \tag{3.6}$$

Suppose there is a finite number of subtrees of T_{max}. Through pruning, a finite sequence of subtrees $T_1, T_2, T_3, ...$ is produced. Each subtree having less terminal nodes.
To find the subtree minimizer for $R_\alpha(T)$ see Breiman et al. (1984).

After pruning, the "right-sized" tree needs to be selected. Cross validation costs is typically used with the minimum cross validation cost being of interest. Often there are a few trees with cross validation costs close to the minimum. Breiman et al. (1984) proposed the "$1 - SE$" rule for selecting the "right-sized" tree. This involves choosing the smallest-sized tree whose cross validation costs do not exceed the minimum cross validation costs plus 1 times the standard error (SE) of the cross validation costs for the minimum cross validation costs tree.

Entropy can be used to assess the goodness of fit where entropy for classification trees which can be defined as:

$$-\sum_{\lambda} \frac{N_\lambda}{N_0} \sum_{\tau} \frac{N_\tau^\lambda}{N_\lambda} log_2 \left(\frac{N_\tau^\lambda}{N_\lambda} \right) \tag{3.7}$$

where λ is a leaf, N_λ is the number of observations on leaf λ, N_0 is the total number of observations in the data set, τ is a level of the response variable, and N_τ^λ is the number of observations on leaf λ that have the response level τ.

3.2 Random forest

A random forest is a collection of tree-structured classifiers. Each tree casts a unit vote for the most popular class. In order for individual trees to not be too correlated, random forests make use of bagging and feature randomness. Bagging is where each individual tree randomly samples from the dataset with replacement. This thus results in different individual trees. When splitting a node, each tree can pick only from a random subset of features (feature randomness). This results in lower correlation and more diversification across the trees (Breiman, 2001).

In selecting a variable in a splitting rule we will explain the Loh method. From the contingency table, this method selects the variable with the smallest $p\text{-}value$ of a chi-square test of association. Let $\boldsymbol{Y}$ denote the target variable. If $\boldsymbol{Y}$ is categorical let it have $\boldsymbol{J}$ categories. Let $\boldsymbol{X}$ denote the input variable and, if $\boldsymbol{X}$ is categorical let it have $\boldsymbol{K}$ categories. So, if both $\boldsymbol{Y}$ and $\boldsymbol{X}$ are categorical a $\boldsymbol{J}X\boldsymbol{K}$ contingency table is produced. If $\boldsymbol{X}$ has interval measurement then, let

$$\boldsymbol{K} = \begin{cases} 3 & \text{if } \boldsymbol{N} < 20\boldsymbol{J}, \\ 4 & \text{otherwise} \end{cases} \tag{3.8}$$

where $\boldsymbol{N}$ is the number of observations in the calculations. Let $\boldsymbol{J} = 2$ if Y is an interval variable. If $\boldsymbol{K} = 3$ then assign $\boldsymbol{X}_i$ to a table column that has the following boundaries:

$$\xi_1 = \bar{\boldsymbol{X}} - \sqrt{3}\hat{\sigma}/3 \tag{3.9}$$

$$\xi_2 = \bar{\boldsymbol{X}} + \sqrt{3}\hat{\sigma}/3 \tag{3.10}$$

Otherwise, use the boundaries:

$$\xi_1 = \bar{\boldsymbol{X}} - \sqrt{3}\hat{\sigma}/3 \tag{3.11}$$

$$\xi_2 = \bar{\boldsymbol{X}} \tag{3.12}$$

$$\xi_3 = \bar{\boldsymbol{X}} + \sqrt{3}\hat{\sigma}/3 \tag{3.13}$$

where $\bar{\boldsymbol{X}}$ is the average of value of $\boldsymbol{X}$ and

$$\sigma^2 = \frac{\sum_i (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^2}{\boldsymbol{N}} \tag{3.14}$$

(Loh, 2002). Other methods for selecting a splitting method include binned search method or the Hothorn, Hornik and Zeileis method. These methods are not described here but can been found in SAS Institute Inc. (2015) or Hothorn et al. (2006).

Once the splitting variable is selected, a splitting rule needs to be defined to know which branch each observation must be split into. As with decision trees, the worth of a split s is the reduction in node impurity (Equation 3.1). The impurity function for the Gini index is given in Equation 3.3 and for the variance reduction is given in Equation 3.4 (Breiman et al., 1984).

In order to assess the validity of a split a centre of each node is computed for both the in-bag and pruning data. In-bag data is part of the split construction and pruning data is part of the split evaluation. A split is pruned when

$$\frac{d(C_{prune}, C_{parent})}{d(C_{inbag}, C_{parent})} < \tau \tag{3.15}$$

where C_{parent} is the in-bag centre of the parent node, C_{inbag} is the in-bag centre of a child node, C_{prune} is the prune centre of a child node, $d(a, b)$ is a measure of distance from a to b, and τ is a fixed number between 0 and 1. For a categorical target, C_ω is the vector of target value proportions and

$$d(a, b) = \sum_{j=1}^{J} (a_j - b_j)(C_{inbag,j} - C_{parent,j}) \tag{3.16}$$

An observation will first be assigned to a single leaf in each decision tree in the random forest. That leaf will be used then to make a prediction depending on the tree that leaf is in. Predictions are averaged over all the trees to predict an observation. For a nominal target, the predicted target category is the category with the largest posterior probability where the posterior probability is the proportion of that category among the bagged training observations in that leaf. If there were to be a tie, the prediction will be the first category that occurs in the training data.

Prediction error is given by average square error which, for a nominal target, is given by

$$ASE = \sum_{i=1}^{N}\sum_{j=1}^{J}\left[\frac{(\delta_{ij} - \hat{p_{ij}})^2}{JN}\right] \tag{3.17}$$

where δ_{ij} equals 1 if the nominal target value j occurs in observation i or equals 0 otherwise, $\hat{p_{ij}}$ is the predicted probability of nominal target value j for observation i, N is the number of observations, J is the number of nominal target values. Misclassification rate can also be assessed for prediction error for a nominal target (Equation 3.28) as well as the log-loss which is defined as

$$LogLoss = -\sum_{i=1}^{N}\sum_{j=1}^{J}\left[\frac{\delta_{ij} log(\ddot{p_{ij}})}{N}\right] \tag{3.18}$$

where $\ddot{p_{ij}}$ is $\hat{p_{ij}}$ truncated away from 0 and 1:

$$\ddot{p_{ij}} = max(min(\hat{p_{ij}}, 1 - 10^{-10}), 10^{-10}). \tag{3.19}$$

Variable importance is the contribution a variable makes to the success of a model where success is simply good prediction of the model. One method of measuring variable importance is loss reduction otherwise known as Gini increase, Gini importance, or impurity reduction. Here, the importance of a variable, say v, is proportional to the sum of the reduction in node impurity, summed over nodes that v splits. For tree T the loss reduction variable importance for input v is

$$I_{loss} \propto \sum_{\omega \epsilon T} 1(v\text{splits}\omega)\Delta Loss(\omega) \tag{3.20}$$

where the sum is over internal nodes ω in T, $1(v\text{splits}\omega)$ is 1 if v is the splitting variable in ω and 0 otherwise. The reduction in loss from splitting ω is $\Delta Loss(\omega)$. A loss function measures how well a model fits data by mapping a response value and a prediction to a number that represents how bad the prediction is. Square error loss is common where

$$\Delta Loss(\omega) = SSE(\omega) - \sum_{b\epsilon B(\omega)} SSE(\omega_b). \tag{3.21}$$

For a categorical target with J classes, square error loss is given by

$$SSE(\omega) = \sum_{i=1}^{N(\omega)} \sum_{j=1}^{J} (\delta_{ij} - \hat{p_j}(\omega))^2 \tag{3.22}$$

where $B(\omega)$ is the set of branches from ω, ω_b is the child node of ω in branch b, $N(\omega)$ is the number of observations in ω, δ_{ij} is 1 if $Y_i = j$ and 0 otherwise,and $\hat{p_j}(\omega)$ is the average δ_{ij} in training data in ω.
For a categorical target, the loss function can also be determined by increasing the margin which is achieved by the probability of the true class minus the maximum probability of the other classes.

$$\Delta Loss(\omega) = SNM(\omega) - \sum_{b\epsilon B(\omega)} SNM(\omega_b). \tag{3.23}$$

Loss reduction variable importance uses the negative of the margin

$$SNM(\omega) = -\sum_{j=1}^{J} N_j(\hat{p_j} - \max_{k\neq j} \hat{p_k}) \tag{3.24}$$

where N_j is the number of class j observations in ω in the data set (Breiman & Cutler, 2003).

Other methods of measuring variable importance include Breiman's method (Breiman, 2001), or Strobl et al. (2008) method.

3.3 Neural networks

A neural network is a series of algorithms that aims to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates (Sarle, 1994).

3.3.1 Bayesian neural network

Traditional neural network lack probabilistic considerations. This can be an issue in applications such as medical diagnosis where representing uncertainty is of critical importance. Bayesian neural networks incorporate weights that are assigned

a probability distribution instead of a single value or point estimate. These probability distributions describe the uncertainty in weights and can be used to estimate uncertainty in predictions.
The graphical model consists of the following two parts:

- $\boldsymbol{G}$ is a directed acyclic graph with nodes representing random variables and arcs between the nodes representing conditional dependency of the random variables.

- $\boldsymbol{P}$ is a set of conditional probability distributions, one for each node conditional on its parents.

Bayesian networks have the following two properties:

- Edges represent "causation"

- Markov property where each node is conditionally independent of its ancestors given its parents.

According to the Markov property, the joint probability distribution of all nodes in the network is given by

$$Pr(G) = Pr(X_1, X_2, ..., X_p) = \prod_{i=1}^{p} Pr(X_i|\pi(X_i)) \tag{3.25}$$

where $\pi(X_i)$ are the parents of node X_i.
In the case where all X_i are discrete variables, the conditional distribution is represented as conditional probability tables. This lists the probability that the child node takes on a certain value for each combination of values of its parents.
In general, a new observation $X = (x_1, x_2, ..., x_p)$ is classified by determining the classification of the target Y that has the largest conditional probability,

$$\arg\max_k Pr(Y = k|x_1, x_2, ..., x_p) \tag{3.26}$$

where

$$Pr(Y = k|x_1, x_2, ..., x_p) \propto Pr(Y = k, x_1, x_2, ..., x_p) = \prod_i Pr(x_i|\pi(X_i))Pr(Y = k|\pi(Y)). \tag{3.27}$$

Different types of Bayesian network classifiers include:

- Naive Bayesian: The target node has a direct edge to each input variable and is the only parent for all nodes. It is assumed that all input variables are conditionally independent of each other given the target.

- Tree-augmented naive Bayesian: The target node has direct edges to each input node and the edges among the input nodes form a tree.
- Bayesian network-augmented naive Bayesian: The target node has a direct edge to each input node and the edges among the input nodes form a Bayesian network.
- Parent-child Bayesian: Input variables can be parents of the target. Edges from the parents of the target to the children of the target and among the children of the target are possible.
- Markov Blanket Bayesian: The Markov blanket includes the target's parents, children and spouses.

In selecting variables, each input variable was tested for conditional independence of the target variable given any other input variable. Only those variables that are conditionally dependent on the target variable given any other input variable are selected.

In learning the tree-augmented naive structure, a maximum spanning tree is constructed. The sum of the weights of all edges is maximum weight among all such tree structures. If there are K variables in the system, the corresponding tree structure will have K nodes and K-1 edges so that all nodes in the graph are connected.

In learning the other Bayesian network types, the following approaches are used:

- The score-based approach: The BIC (Bayesian information criterion) score is used to measure how well a structure fits the data and then finds the structure that has the best score.
- The constraint-based approach: Independence tests (chi-square test or mutual information test) is used to determine the edges and directions among the nodes. The BIC score is used to determine the the directions of thee edges.

The parents of the target is first learnt in the parent-child and Markov blanket structures. Next, parents of the input variables which have the highest BIC with the target variable is learnt. Then, the parents with the next highest BIC is learnt and so on. The edges are determined by independence tests and are oriented by independence tests and BIC score (Liu et al., 2017).

3.4 Goodness of fit

The following metrics can be used to assess goodness of fit for classification methods: entropy, Gini index, misclassification rate, average square error, residual sum of squares, sensitivity, specificity, precision and confusion matrix.

Misclassification Rate

This refers to the number of incorrectly predicted observations and is defined as:

$$Misc = \frac{1}{N_0} \sum \begin{cases} 0 & \text{if prediction is correct,} \\ 1 & \text{otherwise.} \end{cases} \tag{3.28}$$

Confusion matrix

A confusion matrix highlights the type of errors being made by the given classifier. Essentially, the confusion matrix shows where the model gets 'confused' in making predictions. Information on actual values in the rows and predicted values in the columns are represented. We define the following terms with respect to classifying whether an individual has a disease or not:

- True Positive (TP): the number of cases correctly identified for those that have the disease.
- False Positive (FP): the number of cases incorrectly identified for those that have the disease.
- True Negative (TN): the number of cases correctly identified as not having the disease.
- False Negative (FN): the number of cases incorrectly identified as not having the disease.

From the confusion matrix we can determine classifier performance by assessing the following fractions:

- Accuracy measures the proportion of actual positives and negatives that are correctly identified as such.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.29}$$

- Sensitivity measures proportion of actual positives that are correctly identified as such.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.30}$$

- Specificity measures the proportion of actual negatives that are correctly identified as such.

$$Specificity = \frac{TN}{TN + FP} \tag{3.31}$$

- Precision measures the proportion of actual positives out of all those that are predicted to be positive.

$$Precision = \frac{TP}{TP + FP}. \tag{3.32}$$

These measures generally relate to medical decisions as to whether a disease is present or not (Beleites et al., 2013).

3.5 Results from the classification methods

3.5.1 Decision tree

All variables, including significant interaction terms that were determined using a multinomial logistic regression model, were applied to the decision tree, as well as the sample weights. The splitting criterion of entropy and the pruning method of cost-complexity was used. The complexity parameter that yields the minimum average misclassification rate is 0.0011 where the minimum average misclassification rate is 0.302. This corresponds to the 21-leaf subtree. After applying the $1 - SE$ Rule, the ”right-sized” tree is a 16-leaf tree. This resulting decision tree is given in Figure 3.1. The diagram reveals that splitting on 10 of the attributes was sufficient to differentiate the three diabetic statuses. The important variables, in descending order, were age, waist-to-height ratio, salt consumption and waist-to-height ratio, high blood pressure medication, waist circumference, BMI and waist-to-height ratio, BMI, having smoked the previous 24 hours, perception of health, and highest education level.

Classification can be seen in Table 3.1 and an accuracy of 70.0% was obtained. The decision tree is quick to predict pre-diabetics, with a sensitivity for pre-diabetics of 91.7%.

Table 3.1: Confusion matrix

Observed	Predicted			
	Non-diabetic	Pre-diabetic	Diabetic	Total
Non-diabetic	28	618	19	665
Pre-diabetic	0	3698	334	4032
Diabetic	0	822	458	1280
Total	28	5138	811	5977

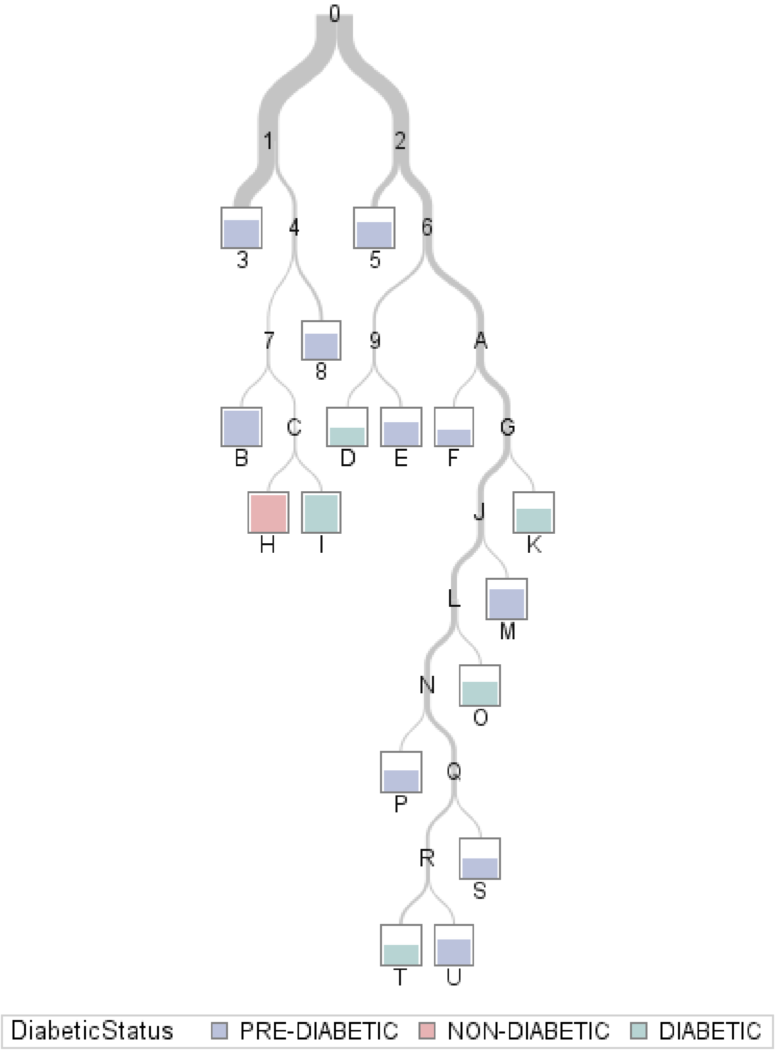

Figure 3.1: Decision tree diagram

3.5.2 Random forest

Again, all the variables of interest including interaction terms were used. The Gini index was used for the split criterion and Loh for the preselection method.
Table 3.2 displays where the random forest gets confused in predicting diabetic status. This method of classification is quick to predict individuals as being pre-diabetic with 94.2% of the population predicted as pre-diabetic. The random forest fails to predict any individuals as being non-diabetic (precision for non-diabetic being 0%). The random forest resulted in an accuracy of 66.6%.
According to the loss reduction variable importance, the 5 most important variables are: taking high blood pressure medication, BMI category, gender, the interaction of salt consumption with waist-to-height ratio, and the interaction of BMI category

Table 3.2: Confusion matrix for the random forest

Observed	Predicted			
	Non-diabetic	Pre-diabetic	Diabetic	Total
Non-diabetic	0	731	11	742
Pre-diabetic	0	4162	130	4292
Diabetic	0	1178	230	1408
Total	0	6071	371	6442

with waist-to-height ratio.

3.5.3 Bayesian neural network

All variables of interest as well as interactions were included in the Bayesian neural network. Our target node being diabetic status. All Bayesian network classifiers were considered. The tree-augmented naive Bayesian network classifier and the Bayesian network-augmented naive Bayesian classifier yielded the highest accuracy (both 76.3%). Results to follow are based on the tree augmented naive Bayesian network classifier.

From Table 3.3, it can been seen that predicted observations are more spread out among the three categories of diabetic status when compared to the confusion matrix of the decision tree and that of the random forest (Table 3.1 & 3.2).

Table 3.3: Confusion matrix for the tree-augmented naive Bayesian network

Observed	Predicted			
	Non-diabetic	Pre-diabetic	Diabetic	Total
Non-diabetic	283	420	39	742
Pre-diabetic	155	3774	363	4292
Diabetic	30	521	857	1408
Total	468	4715	1259	6442

Figure 3.2 shows the generated diagram for the tree-augmented naive Bayesian network classifier applied to our data. The target node has an edge to each input node and the edges among the input nodes form a tree. We have 20 out of 26 variables that were selected into the model.

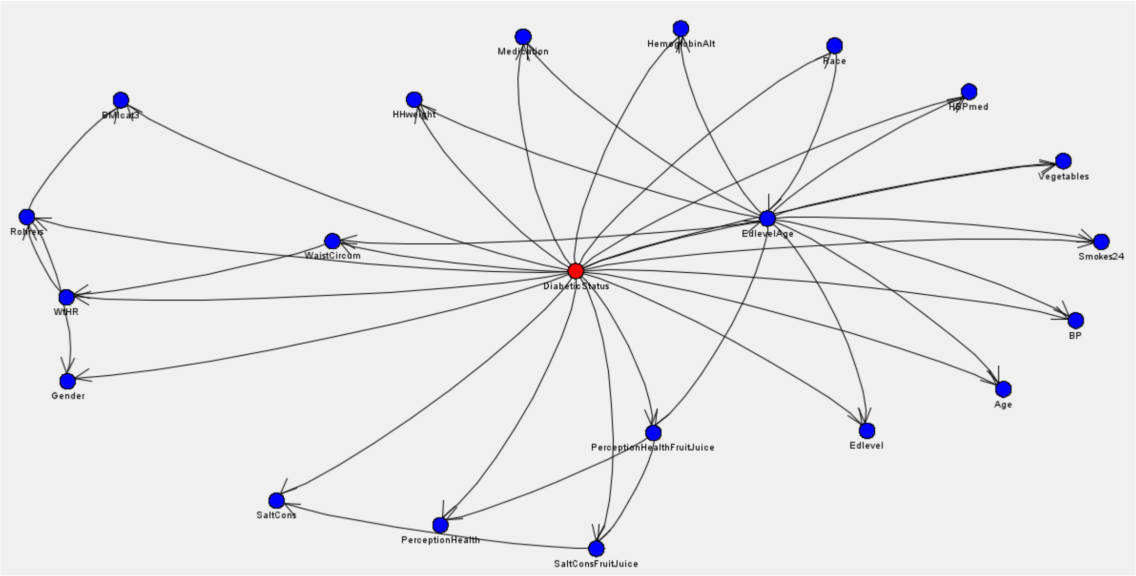

Figure 3.2: Tree-augmented network diagram

Though all the three classification methods provided reasonably good classification and accuracy rates, they failed to account for the complex survey design of the data. These methods also give us an idea of which risk factors are important however, not the significance of the risk factor. In addition, these methods are not able to provide an estimated effect that each variable has on a person's diabetic status. We therefore now move onto regression models, which account for the complex survey design of the data as well as assess which risk factors are significant in classifying diabetic status.

Chapter 4

Generalized Linear Models

Generalized linear models are used in data analysis to describe the relationship between an outcome variable and one or more explanatory variables. Generalized linear models are an extension of the general linear model to address the restrictions on the general linear model. When considering a discrete outcome variable, the logistic regression is most frequently used (Hosmer Jr et al., 2013). A discrete outcome could either be ordinal or nominal in nature. The logistic regression model can be further extended to handle data coming from a complex survey design (Heeringa et al., 2010).

The response variable $Y_i, i = 1, 2, ...n$ follows a distribution that belongs to the exponential family of distributions whose densities can be written in the form

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} \tag{4.1}$$

where θ_i is the canonical parameter, ϕ is the dispersion parameter and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions. The function $a(\phi)$ has the following form $a(\phi) = \phi/w_i$, where w_i known as the prior weight, usually 1. It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{4.2}$$

$$Var(Y_i) = \sigma_i^2 = a(\phi)\, b''(\theta_i) \tag{4.3}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. $b''(\theta_i)$ is a function of the mean and referred to as the variance function, $v(\mu_i)$. When $a(\phi) = \phi$ the variance has the simpler form

$$Var(Y_i) = \sigma_i^2 = \phi v(\mu_i). \tag{4.4}$$

The variance of a generalized linear model is non-constant where it may vary across the responses. When $a(\phi) > 1$ then $Var(Y_i) > v(\mu_i)$ and the model is overdispersed. Similarly, if $a(\phi) < 1$ the model will be underdispersed.
The generalized linear model consists of:

- Linear predictor:
$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

- A link function that describes how the mean $E(Y_i) = \mu_i$ depends on the linear predictor
$$g(\mu_i) = \eta_i$$
where g is a monotone, differentiable function.
The canonical link function is the function that makes the linear predictor $\boldsymbol{\eta}$ the same as the canonical paremeter $\boldsymbol{\theta}$. Therefore, the canonical link function is given by $g(\boldsymbol{\mu}) = \boldsymbol{\theta}$.

- A variance function that describes how the variance $Var(Y_i)$ depends on the mean
$$Var(Y_i) = \phi V(\mu)$$
where the dispersion parameter ϕ is a constant (Nelder & Wedderburn, 1972).

4.1 Parameter estimation

According to Gill (2000), maximum likelihood estimation method is the most popular technique in applied statistics for estimating parameters. Thus, it is no surprise that generalized linear models make use of this technique. The log-likelihood for the ith observation is given by

$$\ell_i = \ln\, f(y_i;\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi). \tag{4.5}$$

We assume that
$$a(\phi) = \frac{\phi}{w_i} \tag{4.6}$$
where w_i are known prior weights.
Since $Y_i, i = 1, 2, ..., n$ are independent, the joint log-likelihood is given by

$$\ell(\boldsymbol{\beta}, \mathbf{y}) = \sum_{i=1}^{n} \ell_i. \tag{4.7}$$

In order to get the maximum likelihood estimate of $\beta_j, j = 0 \dots p$ we need to solve the score equation

$$\frac{\partial \ell}{\partial \beta_j} = 0.$$

To do this, we apply the chain rule

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Using Equation 4.5, we get

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}.$$

Since $\mu_i = b'(\theta_i)$, $Var(Y_i) = a(\phi)v(\mu_i)$ and $\eta_i = \sum_j \beta_j \, x_{ij}$, we have

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \mu_i}{a(\phi)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = b''(\theta_i) = v(\mu_i)$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Thus,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \, \frac{1}{v(\mu_i)} \, \frac{\partial \mu_i}{\partial \eta_i} \, x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \, W_i \, \frac{\partial \eta_i}{\partial \mu_i} \, x_{ij} \end{aligned}$$

where W_i is referred to as the iterative weights given by

$$\begin{aligned} W_i &= \frac{1}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 v_i^{-1} & (4.8) \\ &= \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 & (4.9) \end{aligned}$$

and $v_i = v(\mu_i)$ is the variance function. Since $\eta_i = g(\mu_i)$, $\dfrac{\partial \mu_i}{\partial \eta_i}$ depends on the link function for the model. Therefore, solving for the score equation

$$\sum_{i=1}^{n} (y_i - \mu_i) \, W_i \, \frac{\partial \eta_i}{\partial \mu_i} \, x_{ij} = 0 \tag{4.10}$$

will give the maximum likelihood estimate of $\boldsymbol{\beta}$.

Solution of Equation 4.10 is usually obtained by an iterative weighted least squares method. Newton Raphson and Fisher Scoring iterative equations can be used where the score $\boldsymbol{U}$ is given by the left hand side of Equation 4.10. In matrix and vector form, Equation 4.10 can be rewritten as

$$\boldsymbol{U} = \boldsymbol{X'W\Delta}(\boldsymbol{Y} - \boldsymbol{\mu}) = \mathbf{0} \tag{4.11}$$

where

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \boldsymbol{W} = \text{Diag} \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix}, \boldsymbol{\Delta} = \text{Diag} \begin{bmatrix} \dfrac{\partial \eta_1}{\partial \mu_1} \\ \dfrac{\partial \eta_2}{\partial \mu_2} \\ \vdots \\ \dfrac{\partial \eta_n}{\partial \mu_n} \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}.$$

Thus, the Newton Raphson iterative equation will then be

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1} \boldsymbol{U}^{(t)} \tag{4.12}$$

and the Fisher Score iterative equation

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{\mathcal{I}}^{(t)})^{-1} \boldsymbol{U}^{(t)} \tag{4.13}$$

with information matrix

$$\begin{aligned} \boldsymbol{\mathcal{I}} &= -E(\boldsymbol{H}) && (4.14) \\ &= -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}\right) && (4.15) \\ &= \mathbf{X}'\, \mathbf{W}\, \mathbf{X} && (4.16) \end{aligned}$$

where $\mathbf{W}$ is known as the weight matrix with diagonal elements given in Equation 4.8. Equation 4.13 can also be represented as

$$\boldsymbol{\mathcal{I}}^{(t)}\, \widehat{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\mathcal{I}}^{(t)}\, \widehat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{U}^{(t)}. \tag{4.17}$$

It can be shown that the right hand side of Equation 4.17 can be written as

$$\mathbf{X}'\, \mathbf{W}^{(t)}\, \mathbf{z}^{(t)}$$

where $\mathbf{W}^{(t)}$ is weight matrix evaluated at $\widehat{\boldsymbol{\beta}}^{(t)}$, and $\mathbf{z}^{(t)}$ has the following elements evaluated at $\widehat{\boldsymbol{\beta}}^{(t)}$

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \tag{4.18}$$

This variable z_i is often called the adjusted dependent variable or the working variable. Therefore, we can obtain

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}'\,\mathbf{W}^{(t)}\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{W}^{(t)}\,\mathbf{z}^{(t)}. \tag{4.19}$$

Thus, each iteration step is the result of a weighted least squares regression of the adjusted variable z_i on the predictors x_i with working weight W_i. Fisher scoring can therefore be regarded as iteratively reweighted least squares carried out on a transformed version of the dependent variable.

It follows that the asymptotic variance (also known as the asymptotic covariance) of this estimate of $\boldsymbol{\beta}$ is the inverse of the information matrix given in Equation 4.16 and can be estimated by

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\,\widehat{\mathbf{W}}\,\mathbf{X})^{-1} \tag{4.20}$$

where $\widehat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $\widehat{\boldsymbol{\beta}}$ and depends on the link function of the model. The dispersion parameter ϕ, in function $a(\phi)$ that is used in the calculation of W_i, gets cancelled out of the iteratively reweighted least squares procedure, thus the value of $\widehat{\boldsymbol{\beta}}$ is the same under any value of ϕ. However, the value of ϕ is required for the calculation of the variance of $\widehat{\boldsymbol{\beta}}$, therefore when ϕ is unknown, it can be estimated using a moment estimator (McCulloch et al., 2001), given by

$$\widehat{\phi} = \frac{1}{n-p-1}\sum_{i=1}^{n}\frac{w_i\,(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)} \tag{4.21}$$

where w_i is the weight defined in Equation 4.1 (Eliason, 1993).

4.2 Goodness of fit

After having fitted a model it is of interest to know how accurately the model reflects the true outcome in the data i.e. goodness of fit. A model is unlikely to produce predicted values that match the data perfectly. Goodness of fit are functions of a residual which is simply the difference between the observed and the fitted value (Hosmer Jr et al., 2013). It is of interest for this discrepancy to be as small as possible.

Such a measure that can assess the goodness of fit of a generalized linear model is known as the deviance. Deviance is defined as -2 times the difference in log-likelihood of the fitted model and the saturated model (the model that fits the data perfectly). The scaled deviance is given by

$$D^s = \frac{-2[\ell(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})]}{\phi} \tag{4.22}$$

where $\ell(\widehat{\boldsymbol{\mu}}\phi, \boldsymbol{y})$ is the log-likelihood maximised over $\hat{\beta}$ for a fixed value of the dispersion parameter ϕ, for the fitted model with $p+1$ parameters, and $\ell(\boldsymbol{y}, \phi, \boldsymbol{y})$ is the log-likelihood for the saturated model where the number of parameters equals the number of observations.
If $\phi = 1$, then the deviance is

$$D = -2[\ell(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})]. \tag{4.23}$$

The deviance converges asymptotically to a χ^2 distribution with $n-p-1$ degrees of freedom. Thus, the fitted model is rejected at α level of significance when the calculated deviance is greater than or equal to χ^2_{n-p-1} (Nelder & Wedderburn, 1972).

Pearson's chi-squared statistic is another measure of fit given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)} \tag{4.24}$$

where $v(\widehat{\mu}_i)$ is the estimated variance function for the distribution under consideration. As with the deviance, this statistic also follows the χ^2 distribution with $n-p-1$ degrees of freedom (Pearson, 1900).

4.3 Logistic regression

Multinomial logistic regression is an extension of the binary logistic regression model where the response is nominal with more than two categories. If responses are ordered then an ordinal logistic regression model could be fit such as the proportional odds model. Independent variables can be nominal or continuous.

4.3.1 Ordinal response

The ordinal logit model, ordered logit model or proportional odds model is used when the response variable is ordinal. Ordinal variables are essentially quantitative in nature where each level is of smaller or greater magnitude than another. In our case, diabetic status is ordinal (non-diabetic, pre-diabetic, diabetic) being derived from the continuous variable HbA1c.

For an ordinal response the following logits can be applied:

- Cumulative logit
  For C outcome categories with probabilities $\pi_1, \pi_2, ..., \pi_C$ the cumulative logits are defined as

$$\text{logit}[P(Y \leq j)] = \ln\left[\frac{P(Y \leq j)}{1 - P(y \leq j)}\right] \quad (4.25)$$

$$= \ln\left[\frac{\pi_1 + \pi_2 + ... + \pi_j}{\pi_{j+1} + \pi_{j+2} + ... + \pi_C}\right], j = 1, 2, ..., C - 1. \quad (4.26)$$

- Adjacent-categories logits
  This is the log odds for pairs of adjacent categories

$$\ln\left[\frac{\pi_j}{\pi_{j+1}}\right], j = 1, ..., C - 1. \quad (4.27)$$

  The adjacent-categories logits are defined with conditional probabilities at each of the $C - 1$ cutpoints

$$\text{logit}[P(Y = j|Y = j \text{ or } Y = j{+}1)] = \ln\left[\frac{P(Y = j|Y = j \text{ or } Y = j + 1)}{1 - P(Y = j|Y = j \text{ or } Y = j + 1)}\right]. \quad (4.28)$$

- Continuation-ratio logits
  Defined as:

$$\ln\left[\frac{\pi_j}{\pi_{j+1} + ... + \pi_C}\right], j = 1, 2, ..., C - 1. \quad (4.29)$$

The proportional odds model assumes that the intercepts depend on j while the slopes are all equal. The model is given by

$$\text{logit}[P(Y \leq j|\boldsymbol{x}_i)] = \alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta}, j = 1, 2, ..., C - 1. \quad (4.30)$$

Where α_j is the intercept at category j, β is the slope parameter and x the covariates. The plot of the $C - 1$ cumulative logits against x would thus be a series of parallel lines at the intercepts $\alpha_1, \alpha_2, ..., \alpha_{C-1}$. The model satisfies

$$\ln\left[\frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)}\right] = \beta(x_1 - x_2), \text{for all } j. \quad (4.31)$$

This is the proportional odds assumption. A score test can be carried out to test the proportional odds assumption where:
H_0: Slopes are equal across response functions.
vs.
H_1: Slopes are not equal across response functions i.e. proportional odds assumption is violated.

The proportional odds assumption has been described as anti-conservative and is almost always rejected particularly when there is a large number of explanatory variables (Brant, 1990), the sample size is large, or there is a continuous explanatory variable in the model (Allison, 1999).

In the case that the proportional odds assumption is violated one can move onto a model that falls between the proportional odds model and the more general model. Peterson & Harrell Jr (1990) proposed the following model for the ith observation:

$$\text{logit}[P(Y \leq j)] = \alpha_j + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, j = 1, ..., C - 1. \tag{4.32}$$

Where predictors $\boldsymbol{x}$ have a proportional odds structure and predictors $\boldsymbol{u}$ do not, hence this is the *partial proportional odds model*. For identifiability, one of the $\boldsymbol{\gamma_j}$, say $\boldsymbol{\gamma_1}$, equals 0 (Agresti, 2010).

To determine which predictors have a proportional odds structure and which do not Brant (1990) proposed comparing separate (correlated) fits to the binary logistic models underlying the overall model. According to Brant (1990) the odds ratio should be the same for each ordered dichotomization of the outcome variable given the proportional odds assumption holds.
Williams (2016) states: *"If several variables violate the assumption, then the gologit model (generalized ordered logit or proportional odds model) offers little in the way of parsimony and more widely known techniques such as multinomial logit may be superior."*

Predictive power of explanatory variables

Concordance index is commonly used to assess the predictive power of a model given its explanatory variables. The concordance index estimates whether the probability of the predictions and the outcomes are in agreement. A concordance index of 0.5 corresponds to its expected value from randomly guessing the response. The higher the value, the better the predictive power. So, a value of 1.0 corresponds to perfect prediction (Agresti, 2010).

4.3.2 Nominal response

While ordinal models can be run as nominal models without violating any assumptions, nominal models cannot be run as ordinal models. In our model assume that the response, Y, is nominal with C categories. A reference or baseline category needs to be defined. Let $Y = 1$ be the reference category.

Assume we have p covariates and a constant term, denoted by the vector, $\boldsymbol{x}$, of length $p + 1$ where $x_0 = 1$. The $C - 1$ logit functions can be denoted by

$$\begin{aligned} g_2(\boldsymbol{x}) &= \ln\left[\frac{P(Y=2|\boldsymbol{x})}{P(Y=1|\boldsymbol{x})}\right] && (4.33)\\ &= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + ... + \beta_{2p}x_p && (4.34)\\ &= \boldsymbol{x}'\boldsymbol{\beta}_2 && (4.35)\end{aligned}$$

$$g_3(\boldsymbol{x}) = \ln\left[\frac{P(Y=3|\boldsymbol{x})}{P(Y=1|\boldsymbol{x})}\right] \qquad (4.36)$$

$$\begin{aligned} &= \beta_{30} + \beta_{31}x_1 + \beta_{32}x_2 + ... + \beta_{3p}x_p && (4.37)\\ &= \boldsymbol{x}'\boldsymbol{\beta}_3 && (4.38)\end{aligned}$$

all the way up to

$$\begin{aligned} g_C(\boldsymbol{x}) &= \ln\left[\frac{P(Y=C|\boldsymbol{x})}{P(Y=1|\boldsymbol{x})}\right] && (4.39)\\ &= \beta_{C0} + \beta_{C1}x_1 + \beta_{C2}x_2 + ... + \beta_{Cp}x_p && (4.40)\\ &= \boldsymbol{x}'\boldsymbol{\beta}_C. && (4.41)\end{aligned}$$

The conditional probabilities of each outcome category given the covariate vector are

$$P(Y=1|\boldsymbol{x}) = \frac{1}{1 + e^{g_2(\boldsymbol{x})} + e^{g_3(\boldsymbol{x})} + \ldots + e^{g_C(\boldsymbol{x})}} \qquad (4.42)$$

$$P(Y=2|\boldsymbol{x}) = \frac{e^{g_2(\boldsymbol{x})}}{1 + e^{g_2(\boldsymbol{x})} + e^{g_3(\boldsymbol{x})} + \ldots + e^{g_C(\boldsymbol{x})}} \qquad (4.43)$$

and so on up to

$$P(Y=C|\boldsymbol{x}) = \frac{e^{g_C(\boldsymbol{x})}}{1 + e^{g_2(\boldsymbol{x})} + e^{g_3(\boldsymbol{x})} + \ldots + e^{g_C(\boldsymbol{x})}}. \qquad (4.44)$$

Let $\pi_j(\boldsymbol{x}) = P(Y = j|\boldsymbol{x})$ for $j = 1, 2, ..., C$.
A general expression for the conditional probability in the three category model is

$$P(Y = j|\boldsymbol{x}) = \frac{e^{g_j(\boldsymbol{x})}}{\sum_{k=0}^{C} e^{g_k(\boldsymbol{x})}} \tag{4.45}$$

where the vector $\boldsymbol{\beta}_1 = \mathbf{0}$ and $g_1(\boldsymbol{x}) = 0$.

Now, consider the conditional likelihood function for a sample of n independent observations is

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\pi_1(\boldsymbol{x}_i)^{y_{1i}} \pi_2(\boldsymbol{x}_i)^{y_{2i}} \pi_3(\boldsymbol{x}_i)^{y_{3i}} \ldots \pi_C(\boldsymbol{x}_i)^{y_{Ci}}] \tag{4.46}$$

where $y_{1i}, y_{2i}, \ldots, y_{Ci}$ are created binary variables to indicate group membership of an observation and $\sum_{i=1}^{C} y_{ji} = 1$.
It follows that the log-likelihood function is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{C} y_{ji} g_j(\boldsymbol{x}_i) - \ln(1 + \sum_{j=1}^{C} \exp(g_j(\boldsymbol{x}_i))). \tag{4.47}$$

To get the likelihood equations, the first partial derivatives of $L(\beta)$ with respect to the $2(p + 1)$ unknown parameters are taken. The general form is given by

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_{i=1}^{n} x_{ki}(y_{ji} - \pi_{ji}) \tag{4.48}$$

for $j = 1, 2, \ldots, C$ and $k = 0, 1, 2, ..., p$, with $x_{1i} = 1$ for each subject and $\pi_{ji} = \pi_j(\boldsymbol{x}_i)$. By setting these equations equal to 0 and solving for $\boldsymbol{\beta}$ we obtain the likelihood estimator, $\widehat{\boldsymbol{\beta}}$. An iterative technique, such as Fisher Scoring (Searle et al., 1992) or Newton Raphson method, is required to obtain the estimate.
The general form of the elements in the matrix of second partial derivatives is:

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{jk'}} = -\sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji}(1 - \pi_{ji}) \tag{4.49}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_{jk} \partial \beta_{j'k'}} = \sum_{i=1}^{n} x_{k'i} x_{ki} \pi_{ji} \pi_{j'i} \tag{4.50}$$

for j and $j' = 1, 2, \ldots, C$ and k and $k' = 0, 1, 2, ...p$. From the matrix of second derivatives we can obtain the $2(p + 1)$ by $2(p + 1)$ observed information matrix, $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})$. The elements of $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})$ are the negatives of the values in equations 4.49 and 4.50

evaluated at $\widehat{\boldsymbol{\beta}}$.The relationship of the covariance matrix of the maximum likelihood estimator and the observed information matrix is given by (Hosmer Jr et al., 2013)

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{I}(\widehat{\boldsymbol{\beta}})^{-1}. \tag{4.51}$$

4.4 Survey logistic regression

Researchers apply sample survey methodology to get an accurate view of a large population. Inferences are then made about the population from the sample survey data. For the inferences to be statistically valid, the sample design must be incorporated in the data analysis else, we are likely to obtain bias estimates and misleading standard errors (Nad, 2012). It is common to have categorical outcomes (binary, ordinal and nominal) in survey research and so, logistic regression is often applied to investigate the relationship between categorical response variables and a set of explanatory variables. Survey logistic regression is thus logistic regression applied to survey data. Survey logistic regression methodology differs from ordinary logistic regression methodology in method's used to estimate the model's parameters and variance estimation (Nad, 2012).

For a complex survey design, each observation is presented by a row vector

$$(w_{ij}, \boldsymbol{y}'_{hij}, y_{hij(C)}, \boldsymbol{x}_{hij}) \tag{4.52}$$

where $h = 1, 2, ..., H$ is the stratum number, $i = 1, 2, ..., n_h$ is the cluster number within stratum h, $j = 1, 2, ..., m_{hi}$ is the unit number within cluster i of stratum h, w_{hij} is the sampling weight, $\boldsymbol{y}_{hij}$ is a $(c-1)$-dimensional column vector of indicator variables. The cth row of the vector is one if the response of the jth member of the ith cluster in stratum h falls in category c, where $c = 1, 2, ..., C-1$. The remaining elements of the vector are zero. $y_{hij}(C)$ is the indicator variable for the (C) category of variable Y. $\boldsymbol{x}_{hij}$ is the k-dimensional row vector of explanatory variables for the jth member of the ith cluster in stratum h. If there is an intercept, then $x_{hij1} \equiv 1$. $\tilde{n} = \sum_{h=1}^{H} n_h$ is the total number of clusters in the entire sample and $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total sample size.

Let $\boldsymbol{\pi_{hij}}$ be the expected vector of the response variable. The pseudo log likelihood is given by

$$L(\boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}((\ln(\boldsymbol{\pi}_{hij}))'\boldsymbol{y}_{hij} + \ln(\pi_{hij(C)}))y_{hij(C)} \tag{4.53}$$

As with logistic regression, the maximum likelihood estimator, $\widehat{\boldsymbol{\beta}}$, is obtained by solving for an iterative equation with the Fisher scoring or Newton Raphson technique (see Anthony (2002)).

Survey logistic regression differs from logistic regression in variance estimation. The complex sample design is taken into account in Taylor expansion approximation for variance estimation. Variances within each stratum are computed and pooled together. When considering Taylor approximation, the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{V}(\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{G}}\widehat{\boldsymbol{Q}}^{-1} \tag{4.54}$$

where

$$\widehat{\boldsymbol{Q}} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m} w_{hij}\widehat{\boldsymbol{D}}_{hij}(\text{diag}(\widehat{\boldsymbol{\pi}}_{hij}) - \widehat{\boldsymbol{\pi}}_{hij}\widehat{\boldsymbol{\pi}}'_{hij})^{-1}\widehat{\boldsymbol{D}}'_{hij} \tag{4.55}$$

$$\widehat{\boldsymbol{G}} = \frac{n-1}{n-p}\sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1} n_h(\boldsymbol{e}_{hi.} - \bar{\boldsymbol{e}}_{h..})'(\boldsymbol{e}_{hi.} - \bar{\boldsymbol{e}}_{h..}) \tag{4.56}$$

$$\boldsymbol{e}_{hi.} = \sum_{j=1}^{m_{hi}} w_{hij}\widehat{\boldsymbol{D}}_{hij}(\text{diag}(\widehat{\boldsymbol{\pi}}_{hij}) - \widehat{\boldsymbol{\pi}}_{hij}\widehat{\boldsymbol{\pi}}'_{hij})^{-1}(\boldsymbol{y}_{hij} - \widehat{\boldsymbol{\pi}}_{hij}) \tag{4.57}$$

$$\bar{\boldsymbol{e}}_{h..} = \frac{1}{n_h}\sum_{i=1}^{n_h}\boldsymbol{e}_{hi.} \tag{4.58}$$

and $\widehat{\boldsymbol{D}}_{\boldsymbol{hij}}$ is the matrix of partial derivatives of the link function with respect to $\boldsymbol{\beta}$ evaluated at $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\pi}}_{hij}$ is also evaluated at $\widehat{\boldsymbol{\beta}}$ (Anthony, 2002).

4.5 Results from the logistic models

4.5.1 Proportional odds model and partial proportional odds model

The survey logistic regression model was initially considered in order to account for the complex survey design of the data. The household sampling weights could thus be considered as well as the strata and cluster level. The ordinal nature of the response level was considered by applying the cumulative logit link. The score test for proportional odds assumption yielded a Chi-square value of 732.8 (p-value<0.0001) which suggested that the model was not a good fit. The AIC was 10045.6 and the model had an accuracy of 68.1%. The concordance index was 64.2% and thus the model was low in predictive power. This model's results can be found in Table 4.1 where the estimate, odds ratio (OR) and 95% confidence interval (CI) is given.

Table 4.1: Proportional odds model regression results

Variable	Estimate	OR (95% CI)
Gender (ref=male)		
Female	-0.0280	0.972 (0.802-1.179)
Race (ref=other)		
Black/African	0.1373	1.147 (0.897-1.466)
Age	0.0118*	1.012 (1.007-1.017)
Highest education level (ref=no education)		
Primary	0.00222*	1.002 (0.803-1.250)
Secondary	-0.1552*	0.856 (0.670-1.094)
Higher	-0.1649*	0.848 (0.615-1.169)
BMI category (ref=underweight)		
Normal	-0.2712	0.762 (0.563-1.032)
Overweight to obese	-0.3049	0.737 (0.535-1.016)
Rohrers' Index	-0.0101	0.990 (0.959-1.022)
Waist circumference	-0.03977	0.996 (0.979-1.013)
Waist-to-height ratio	4.4620*	86.664 (3.581->999.99
Haemoglobin level adjusted for altitude and smoking	0.00459	1.005 (0.967-1.043)
Blood pressure (ref=normal)		
Abnormal	0.0356	1.036 (0.903-1.189)
Taking high blood pressure medication (ref=no)		
Yes	0.2972*	1.346 (1.110-1.632)
Taking Medication (ref=no)		
Yes	0.0758	1.079 (0.903-1.288)
Household's perception of health (ref=good to excellent)		
Poor to average	0.0878	1.092 (0.961-1.240)
Household's approach towards salt consumption (ref=positive)		
Negative	0.0255	1.026 (0.892-1.179)
Household's consumption of processed foods	-0.0851*	0.918 (0.861-0.980)
Household's consumption of fruit the previous day (ref=no)		
Yes	0.00558	1.006 (0.883-1.145)

Table 4.1 – *Continued from previous page*

Variable	**Estimate**	**OR (95% CI)**
Household's consumption of vegetables the previous day (ref=no)		
Yes	0.0672	1.069 (0.941-1.216)
Household's consumption of sugary drinks the previous day (ref=no)		
Yes	-0.00255	0.997 (0.860-1.157)
Household's consumption of fruit juice the previous day (ref=no)		
Yes	0.1207	1.128 (0.921-1.383)
Smoking the previous 24hrs (ref=no)		
Yes	-0.1266	0.881 (0.719-1.080)
Wealth index factor score combined	0.0508	1.052 (0.969-1.142)

*significant at 5% level of significance

Next, a partial proportional odds model was of interest to account for certain variables not following the proportional odds assumption, which were violated in the proportional odds model. In order to determine those variables that satisfy the proportional odds assumption, a linear hypothesis test was performed for each variable in the non-proportional odds model. A variable that was significant at a relaxed p-value of 10% did not satisfy the proportional odds assumption. Variables that did satisfy the proportional odds assumption were gender and having had fruit juice the previous day. Results of the partial proportional odds model can be found in Tables 4.2 and 4.3.

Table 4.2: Partial proportional odds model regression results for pre-diabetic vs non-diabetic

Variable	**Estimate**	**OR (95% CI)**
Gender (ref=male)		
Female	-0.2328	0.792 (0.620-1.012)
Race (ref=other)		
Black/African	0.4415*	1.1555 (1.201-2.013)
Age	-0.0198*	0.980 (0.974-0.987)
Highest education level (ref=no education)		
Primary	-0.2431	0.784 (0.526-1.169)
Secondary	-0.2681	0.765 (0.517-1.131)

Table 4.2 – *Continued from previous page*

Variable	Estimate	OR (95% CI)
Higher	-0.2526	0.777 (0.478-1.262)
BMI category (ref=underweight)		
Normal	-0.1115	0.895 (0.663-1.207)
Overweight to obese	-0.1604	0.852 (0.551-1.316)
Rohrers' Index	-0.0857*	0.918 (0.871-0.967)
Waist circumference	-0.0351*	0.965 (0.941-1.990)
Waist-to-height ratio	6.4640*	641.605 (6.991->999.99
Haemoglobin level adjusted for altitude and smoking	14.9388*	1.092 (1.044-1.142)
Blood pressure (ref=normal)		
Abnormal	-0.0849	0.919 (0.760-1.110)
Taking high blood pressure medication (ref=no)		
Yes	-0.0847	0.919 (0.675-1.250)
Taking Medication (ref=no)		
Yes	-0.1035	0.902 (0.676-1.202)
Household's perception of health (ref=good to excellent)		
Poor to average	-0.0439	0.957 (0.812-1.129)
Household's approach towards salt consumption (ref=positive)		
Negative	0.0101	1.010 (0.853-1.197)
Household's consumption of processed foods	-0.0415	0.959 (0.873-1.054)
Household's consumption of fruit the previous day (ref=no)		
Yes	-0.0920	0.912 (0.771-1.079)
Household's consumption of vegetables the previous day (ref=no)		
Yes	-0.0311	0.969 (0.820-1.146)
Household's consumption of sugary drinks the previous day (ref=no)		
Yes	0.00772	1.008 (0.848-1.198)
Household's consumption of fruit juice the previous day (ref=no)		
Yes	0.2529*	1.288 (1.081-1.630)
Smoking the previous 24hrs (ref=no)		
Yes	0.1910	1.210 (0.971-1.509)

Table 4.2 – *Continued from previous page*

Variable	**Estimate**	**OR (95% CI)**
Wealth index factor score combined	0.0851	1.089 (0.989-1.198)

*significant at 5% level of significance

Table 4.3: Partial proportional odds model regression results for diabetic vs pre -diabetic

Variable	**Estimate**	**OR (95% CI)**
Gender (ref=male)		
Female	-0.0889	0.915 (0.748-1.119)
Race (ref=other)		
Black/African	0.2129	1.237 (0.987-1.551)
Age	-0.0246*	0.976 (0.971-0.981)
Highest education level (ref=no education)		
Primary	-0.1165	0.890 (0.703-1.128)
Secondary	-0.0723	0.930 (0.727-1.190)
Higher	0.0485	1.050 (0.744-1.481)
BMI category (ref=underweight)		
Normal	-0.0325	0.968 (0.658-1.424)
Overweight to obese	-0.0855	0.918 (0.600-1.404)
Rohrers' Index	-0.0105	0.990 (0.959-1.021)
Waist circumference	-0.0225*	0.978 (0.961-0.995)
Waist-to-height ratio	-1.0992	0.333 (0.014-8.117)
Haemoglobin level adjusted for altitude and smoking	0.0576*	1.059 (1.023-1.097)
Blood pressure (ref=normal)		
Abnormal	-0.0880	0.916 (0.797-1.052)
Taking high blood pressure medication (ref=no)		
Yes	-0.2767*	0.758 (0.628-0.916)
Taking Medication (ref=no)		
Yes	-0.1048	0.901 (0.751-1.079)
Household's perception of health (ref=good to excellent)		
Poor to average	-0.0810	0.922 (0.805-1.056)

Table 4.3 – *Continued from previous page*

Variable	**Estimate**	**OR (95% CI)**
Household's approach towards salt consumption (ref=positive)		
Negative	0.0170	1.017 (0.881-1.174)
Household's consumption of processed foods	0.0726*	1.075 (1.001-1.155)
Household's consumption of fruit the previous day (ref=no)		
Yes	0.00905	1.009 (0.880-1.158)
Household's consumption of vegetables the previous day (ref=no)		
Yes	-0.0978	0.907 (0.789-1.042)
Household's consumption of sugary drinks the previous day (ref=no)		
Yes	-0.0328	0.968 (0.838-1.117)
Household's consumption of fruit juice the previous day (ref=no)		
Yes	-0.1285	0.879 (0.721-1.072)
Smoking the previous 24hrs (ref=no)		
Yes	0.1940	1.214 (0.974-1.514)
Wealth index factor score combined	0.00996	1.010 (0.932-1.094)

*significant at 5% level of significance

To further examine the the proportional odds assumption, we fitted two separate binary logistic regression models for each dichotomised response (non-diabetic versus pre-diabetic and diabetic versus pre-diabetic). These two regression models indicated different significant variables. For the first binary logistic regression model of non-diabetic versus pre-diabetic, the significant variables were race, age, Rohrers' index, waist circumference, waist-to-height ratio, haemoglobin level and fruit juice. For the second binary logistic regression model of diabetic versus pre-diabetic, the significant variables were race, age, waist circumference, haemoglobin level, high blood pressure medication being taken and consumption of processed foods. The variables that do not correlate between the two models (Rohrers' index, waist-to-height ratio, fruit juice, high blood pressure medication and consumption of processed foods) further suggest that the proportional odds model is not a good fit. Furthermore, the significance of fruit juice only in the first binary logistic regression model violates the finding from the linear hypothesis test that this variable satisfies the proportional odds assumption.

When comparing the results of the proportional odds model and the two binary logistic regression models, violation of the proportional odds assumption is apparent. The ordinal beta coefficient for waist-to-height ratio in the proportional odds model is 4.4620. Taking the ordinal beta coefficient for waist-to-height ratio (6.4640) in the logistic regression model of non-diabetic vs pre-diabetic, this model overestimates the impact of waist-to-height ratio. Whereas, the model of diabetic vs pre-diabetic underestimates the impact of waist-to-height ratio (beta coefficient of -1.0992). The use of an ordered logit model when its assumptions are violated creates a misleading impression of how the outcome and explanatory variables are related.

4.5.2 Final survey logistic regression model

Finally, a multinomial survey logistic regression model was fitted. Diabetic status, non-diabetic and diabetic, were each modelled against pre-diabetic. Interaction effects were then considered. Every pair of variables was considered as interaction effects. Each interaction was entered into the main model one at a time with the main model effects alone and no other interaction term. The interaction's significance (p-$value$) and the model's Akaike information criterion (AIC) were recorded. The interaction term that produced the lowest AIC was then put into the main model and one by one the next four interaction terms that produced the lowest AIC were considered. The interaction term salt consumption and waist-to-height ratio yielded the lowest AIC and was thus permanently entered into the model. The next four interaction terms were entered back into the model one at a time and their p-$value$ and the model's AIC again recorded. Highest education level and age yielded the lowest AIC with interaction terms being significant and was then permanently added into the model. The same process followed with BMI and waist-to-height ratio, perception of health and fruit juice consumed, and salt consumption and fruit juice, all yielding a lower AIC and thus permanently entered into the model. The main model thus consists of 5 interaction effects and the resulting AIC was 10045.594.

Results from the multinomial survey logistic regression model can be found in Table 4.4. A separate model for non-diabetic against pre-diabetic and diabetic against pre-diabetic was formed. First, consider the multinomial survey logistic regression of non-diabetics against pre-diabetics. The odds of a female being non-diabetic rather than pre-diabetic is 1.336 times that of males. For a unit increase in Rohrers' Index, the odds of being non-diabetic is 0.911 times that of being pre-diabetic. For a unit increase in waist circumference, the odds of being non-diabetic is 0.958 times that of being pre-diabetic. For a unit increase in haemoglobin level, the odds of being non-diabetic is 1.110 that of being pre-diabetic. For a unit increase in wealth index score factor, the odds of being non-diabetic is 1.138 times that of being pre-diabetic. Note that the odds ratios and their 95% confidence intervals for variables involved in the interaction terms are not displayed in Table 4.4 below as they are not meaningful to be interpreted on their own.

Table 4.4: Multinomial survey logistic regression results

Variable	Non-diabetic		Diabetic	
	Estimate	**OR (95%CI)**	**Estimate**	**OR (95%CI)**
Gender (ref=male)				
Female	0.2899*	1.336 (1.030-1.734)	-0.05390	0.948 (0.766-1.172)
Race (ref=other)				
Black/African	-0.2497	0.779 (0.584-1.039)	0.3073*	1.360 (1.063-1.738)
Age	-0.04878*		-0.00032	
Highest education level (ref=no education)				
Primary	-1.8625*		-1.5298*	
Secondary	-1.8561*		-1.9035*	
Higher	-1.5278		-0.8961	
BMI category (ref=underweight)				
Normal	-1.7590		-5.4770*	
Overweight to obese	-1.3796		-7.2601*	
Rohrers' Index	-0.09354*	0.911 (0.857-0.967)	-0.03442	0.966 (0.931-1.002)
Waist circumference	-0.04314*	0.958 (0.932-0.984)	0.005697	1.006 (0.987-1.025)
Waist-to-height ratio	3.3204		-11.5736*	
Haemoglobin level	0.1046*	1.110 (1.058-1.165)	-0.04185*	0.959 (0.924-0.996)
Blood pressure (ref=normal)				
Abnormal	-0.2026	0.817 (0.664-1.004)	0.09351	1.098 (0.948-1.272)
Taking high blood pressure medication (ref=no)				
Yes	-0.1126	0.893 (0.649-1.230)	0.2474*	1.281 (1.052-1.560)
Taking Medication (ref=no)				
Yes	-0.2118	0.809 (0.601-1.090)	0.1218	1.129 (0.936-1.363)
Household's perception of health (ref=good to excellent)				
Poor to average	0.002898		0.04444	
Household's approach towards salt consumption (ref=positive)				

Table 4.4 – *Continued from previous page*

Variable	Non-diabetic		Diabetic	
	Estimate	**OR (95%CI)**	**Estimate**	**OR (95%CI)**
Negative	-2.3892*		-0.6598	
Processed foods	-0.08509	0.981 (0.829-1.017)	-0.07581	0.927 (0.857-1.002)
Household's consumption of fruit the previous day (ref=no)				
Yes	-0.00413		-0.02819	
Household's consumption of vegetables the previous day (ref=no)				
Yes	-0.01871		0.1236	
Household's consumption of sugary drinks the previous day (ref=no)				
Yes	-0.1034		0.02797	
Household's consumption of fruit juice the previous day (ref=no)				
Yes	0.09669		-0.3394*	
Smoking the previous 24hrs (ref=no)				
Yes	0.2044		-0.1941	
Wealth index factor score	0.1296*	1.138 (1.029-1.259)	-0.01392	0.986 (0.907-1.072)
Interaction terms				
Waist-to-height ratio & Salt consumption				
Negative	4.6544*		1.1367	
Positive (ref)				
Age & Highest education level				
Primary	0.03240*	0.984 (0.971-0.997)	0.02563*	1.026 (1.016-1.035)
Secondary	0.03359*	0.985 (0.976-0.994)	0.03130*	1.031 (1.025-1.038)
Higher	0.02529	0.977 (0.956-0.998)	0.008625	1.008 (0.993-1.024)
No education (ref)				
Waist-to-height ratio & BMI				
Normal	3.7008		13.2402*	
Overweight to obese	3.0405		17.0990*	
Underweight (ref)				
Perception of health & Fruit juice				
Poor to average, Yes	-0.6577*		0.6341*	

Table 4.4 – *Continued from previous page*

Variable	**Non-diabetic**		**Diabetic**	
	Estimate	**OR (95%CI)**	**Estimate**	**OR (95%CI)**
Good to excellent, No (ref)				
Salt Consumption & Fruit juice				
Negative, Yes	0.8404*		0.7212*	
Positive, No (ref)				

*significant at 5% level of significance

4.5.3 Confusion Matrix

The individual probabilities for each response level from the multinomial survey logistic regression model were considered. The largest individual predicted probability was then noted as the predicted response for the model. A confusion matrix is shown in Table 4.5.

Table 4.5: Confusion matrix for the multinomial survey logistic regression model

Observed	Predicted			Total
	Non-diabetic	Pre-diabetic	Diabetic	
Non-diabetic	0	727	15	742
Pre-diabetic	1	4057	234	4292
Diabetic	2	1044	362	1408
Total	3	5828	611	6442

From the above table (Table 4.5), it was calculated that 68.6% of the observations were correctly classified by the model. The sensitivity for being pre-diabetic is 94.5%, 25.7% for diabetic and 0% for non-diabetic. The model is thus poor at finding relevant instances of being non-diabetic. This can also been seen by the precision of being classified as non-diabetic which is 0. All the observed non-diabetics were classified as pre-diabetic (98.0%) or diabetic (2.0%). Our concern though, is primarily the predicted classes of the observed diabetics. It is of importance that the observed diabetics do not get classified as a lower class (pre-diabetic and more importantly, non-diabetic).

Chapter 5

Spatial Linear Mixed Model

It is important to account for spatial autocorrelation in a model else, the model's interpretability is unreliable (Cressie, 1992). The presence of spatial autocorrelation in the residuals of a model can be accounted for by including the effects of longitude and latitude in a model (Haining, 2003). We present here how to include the effects of longitude and latitude in a mixed model to account for the spatial autocorrelation observed in the previous chapter.

5.1 Spatial autocorrelation with results

Spatial autocorrelation can be defined as the degree to which one object is similar to other nearby objects. Strongly correlated data reduces the statistical power of inference making a model untrustworthy. Given a set of features and an associated attribute, Moran's Index (I) evaluates whether the pattern expressed is clustered, dispersed or random. Moran's I statistic for spatial autocorrelation is given by

$$I = \frac{n \sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{W \sum_{i=1}^{n}(z_i - \bar{z})} \tag{5.1}$$

where w_{ij} is the spatial weight between feature i and j, $W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ and z_i is the location of an attribute for feature *i* with mean $\bar{z}$ (Cliff & Ord, 1972).

The sampling design and the sampling weights are well accounted for in the survey logistic model. Accordingly, the goodness of fit as well as all the model diagnostics show the model is an acceptable fit. However, the spatial plot of the residuals from this model (Figures 5.1 and 5.2) indicate pockets of spatial residual clusters. Both the non-diabetic set and the diabetic set were found to be significantly spatially autocorrelated (Moran's I: 0.014068 and 0.042885, p-value: 0.000517 and 0.019463, respectively). Here, Anselin Local Moran's I for cluster and outlier analysis was applied to both the non-diabetic and diabetic set of mean residuals. From Figure 5.1, it can be seen that in the Northern Cape, North West and Gauteng provinces, high-high clusters of non-diabetics can be seen as well as low-high outliers among these. While in the Eastern Cape and Free State, low-low clusters of non-diabetics and high-low outliers are observed. From Figure 5.2, low-low clusters of diabetics are found in Gauteng and the North West province while high-high clusters of diabetics are observed in Limpopo. High-low outliers can be seen in Gauteng, the Eastern Cape and the Western Cape.

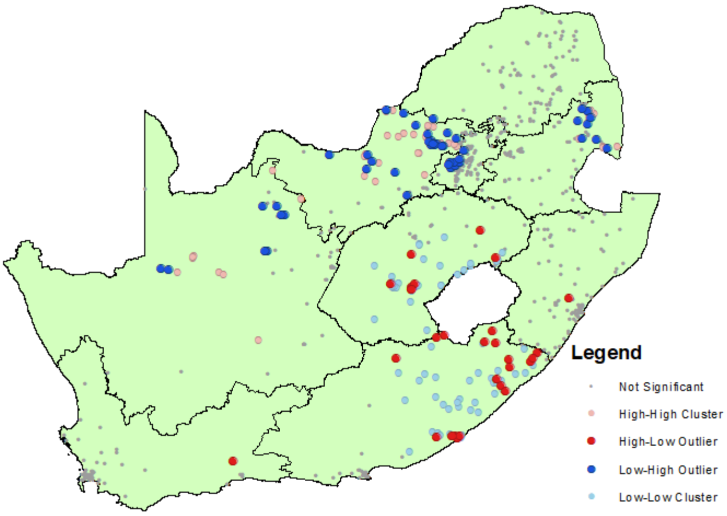

Figure 5.1: Anselin local Moran's I applied to the mixed model for non-diabetics

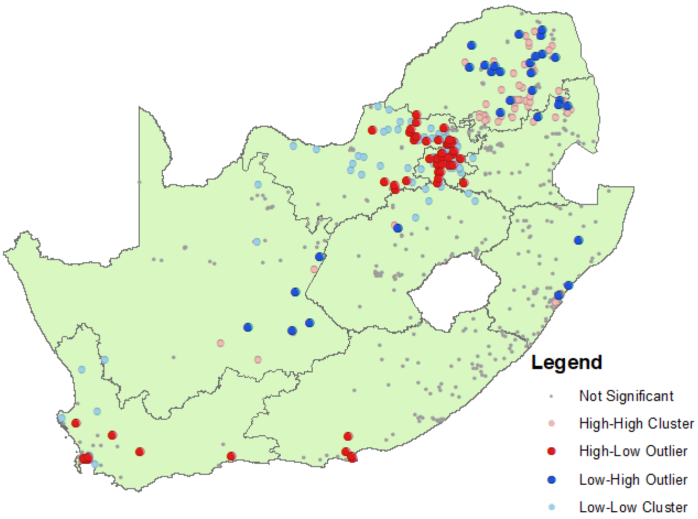

Figure 5.2: Anselin local Moran's I applied to the mixed model for diabetics

Based on the presence of significant autocorrelation in the residuals, it is therefore necessary to account for this spatial autocorrelation in order to obtain trust worthy results. Ways in which to do so will be discussed in the next chapter, specifically ways in which to include the effects of longitude and latitude.

5.2 Generalized mixed models

A generalized linear mixed model allows response variables to come from distributions (namely, the exponential family of distributions) other than the normal distribution. For this to be possible a link function is required. Let the linear predictor, $\boldsymbol{\eta}$, be the combination of fixed and random effects excluding the residuals

$$\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}. \tag{5.2}$$

The link function, $g(.)$, relates the conditional mean outcome $\boldsymbol{y}$ to the linear predictor

$$g(E(\boldsymbol{y}|\gamma)) = \boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} \tag{5.3}$$

where

- $\boldsymbol{y}$ is an $n \times 1$ vector of response variables, where $\sum\limits_{i=1}^{m} n_i = n$ is the total number of observations and m is the number of values that $\boldsymbol{y}$ takes.
- $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix for the fixed-effects.
- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of fixed effect coefficients.
- $\mathbf{Z}$ is the $n \times q$ design matrix for the random effects.
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effect coefficients.

We assume that, for an $n \times 1$ vector of errors, $\boldsymbol{\epsilon}$, and the random effects in $\boldsymbol{\gamma}$

$$\boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{R}_{n\times n})$$

$$\gamma \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G}_{q\times q})$$

and

$$Cov(\boldsymbol{\epsilon}, \boldsymbol{\gamma}) = \boldsymbol{0}_{n\times q}.$$

Therefore, for $\boldsymbol{y}|\gamma$ a member of the exponential family of distributions, it follows that

$$E(\boldsymbol{y}|\gamma) = \mu = g^{-1}(\boldsymbol{X\beta} + \boldsymbol{Z\gamma})$$

$$Var(\gamma) = \boldsymbol{G}$$

$$Var(\boldsymbol{y}|\gamma) = \boldsymbol{A\mu}^{1/2}\boldsymbol{RA\mu}^{1/2}$$

$$Var(\boldsymbol{y}) = \boldsymbol{A}^{1/2}\boldsymbol{RA}^{1/2}$$

where $\boldsymbol{A\mu}$ is a diagonal matrix containing variance functions since $Var(\boldsymbol{y}) = \phi v(\boldsymbol{\mu})$ for distributions belonging to the exponential family of distributions. $\boldsymbol{G}$ is the variance-covariance matrix of the random effects, $\boldsymbol{R}$. It is square, symmetric and positive semidefinite. $\boldsymbol{R}$ is the error term variance-covariance matrix which includes the spatial correlation (McCulloch et al., 2001).

There are many methods to estimating the unknown parameters. The method of least squares is commonly used for normally distributed linear models however, proves to be problematic for other distributions. Thus, only the maximum likelihood estimation method will be outlined.

Issues in estimation for generalized linear mixed models exist, namely obtaining the marginal log-likelihood function by integrating out the random effects from the joint distribution. Numerical integration is only successful when the number of random effects is small. The contribution of the i^{th} cluster to the likelihood is given by

$$f_i(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i\,|\,\boldsymbol{G})\, d\boldsymbol{\gamma}_i \tag{5.4}$$

where $f(\boldsymbol{\gamma}_i\,|\,\boldsymbol{G})$ is the distribution of the random effects.
Therefore, the complete likelihood function for $\boldsymbol{\beta}, \boldsymbol{G}$ and ϕ is given by

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{G}, \phi) &= \prod_{i=1}^{m} f_i(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) \\ &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i\,|\,\boldsymbol{G})\, d\boldsymbol{\gamma}_i. \end{aligned} \tag{5.5}$$

From here we require approximations to evaluate the likelihood function given in Equation 5.5. We explain here Laplace approximation. This method involves approximating the integrand itself (Jiang, 2007). Suppose one wishes to approximate an integral in the form

$$\int e^{Q(\boldsymbol{x})} dx \tag{5.6}$$

where $Q(\boldsymbol{x})$ is a known, unimodal function, and $\boldsymbol{x}$ is a $q \times 1$ vector of variables and achieves its minimum value at $x = \tilde{x}$ with $Q'(\tilde{x}) = 0$ and $Q''(\tilde{x}) < 0$. Then, by Taylor expansion

$$Q(\boldsymbol{x}) \approx Q(\tilde{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x} - \tilde{\boldsymbol{x}})'Q''(\tilde{\boldsymbol{x}})(\boldsymbol{x} - \tilde{\boldsymbol{x}}) \tag{5.7}$$

where $Q''(\tilde{\boldsymbol{x}})$ is the Hessian of Q evaluated at $\tilde{\boldsymbol{x}}$.

This yields an approximation to Equation 5.6:

$$\int e^{Q(\boldsymbol{x})}dx \approx (2\pi)^{\frac{q}{2}}|Q''(\tilde{\boldsymbol{x}})|^{-\frac{1}{2}}\, e^{-Q'(\tilde{\boldsymbol{x}})}. \tag{5.8}$$

Other methods of estimation include penalized quasi-likelihood, tests of zero variance components, maximum hierarchical likelihood estimation (Jiang, 2007).

5.3 Generalized additive mixed models

Unlike the generalized linear model, the generalized additive mixed model are flexible statistical methods allowing nonlinear effects in the model. Generalized additive models take the following form

$$g(\mu) = \eta = \alpha + f_1(x_1) + f_2(x_2) + ... + f_p(x_p) \tag{5.9}$$

where the mean μ of the response is related to $g(\mu) = \eta$ which is a link function, $f_1(x_1), f_2(x_2), ..., f_p(x_p)$ are smooth functions making up the additive component (Hastie & Tibshirani, 1990).
Generalized additive mixed models can be extended from generalized linear mixed models Chen (2000):

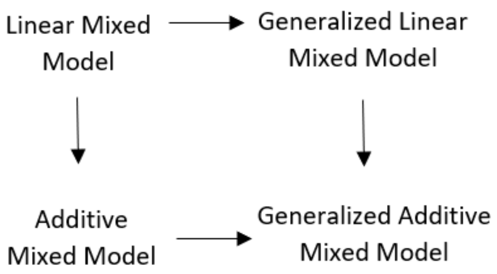
Linear Mixed Model
Generalized Linear Mixed Model
Additive Mixed Model
Generalized Additive Mixed Model

Figure 5.3: Diagram of mixed model extensions by Chen (2000)

The nonlinear form can be selected from several non-parametric smooth functions. Here, we will focus on splines specifically, B-splines or ″basis splines″.

B splines take the following form

$$f(x) = \sum_{k=1}^{K} \xi_k B_k(x) \tag{5.10}$$

where $B_k(x)$ is the k^{th} B-spline basis function of degree d over the domain $[a, b]$. A B-spline includes all polynomials of the same degree or less.

5.4 Results from the spatial model

Upon finding the residuals of the survey logistic regression model to be spatially autocorrelated in the 5.1, we go on to include the effects of longitude and latitude in the model. These are based on the geographical coordinates of the clusters selected in the SADHS.

Using the separate models discussed in 5.1, we considered fitting spatial covariance structure models based on longitude and latitude. No change in regression estimates were seen with the different spatial covariance structures. We utilised an exponential spatial covariance structure (with longitude and latitude as coordinates) for both models. Next, the different estimation methods were considered. Since, the Laplace method resulted in the most number of significant variables for both models, it was utilised. Mean residuals from each model were again obtained and Moran's I was used to check for spatial autocorrelation. Still, both models, were found to be spatially autocorrelated.

Next, we focused on surface correlation by adding longitude and latitude as fixed spline effects into the model to account for the non-linear nature of the coordinates. We utilised B-spline basis for longitude and latitude separately. All other effects were considered to be linear. Again, Moran's I was used to check for spatial autocorrelation. The residuals of each of the two models were no longer found to be spatially autocorrelated, where Moran's I was -0.003071 (p-value=0.716986) for the non-diabetic set and Moran's Index was 0.000685 (p-value=0.634226) for the diabetic set. Thus, the addition of the fixed spline effects of longitude and latitude was sufficient to account for spatial autocorrelation in the models.

Again, Anselin local Moran's I was applied to the two generalized additive models. The impact of the addition of the spline effects of longitude and latitude can be seen by less significant clustering in Figures 5.4 and 5.5 compared to Figures 5.1 and 5.2.

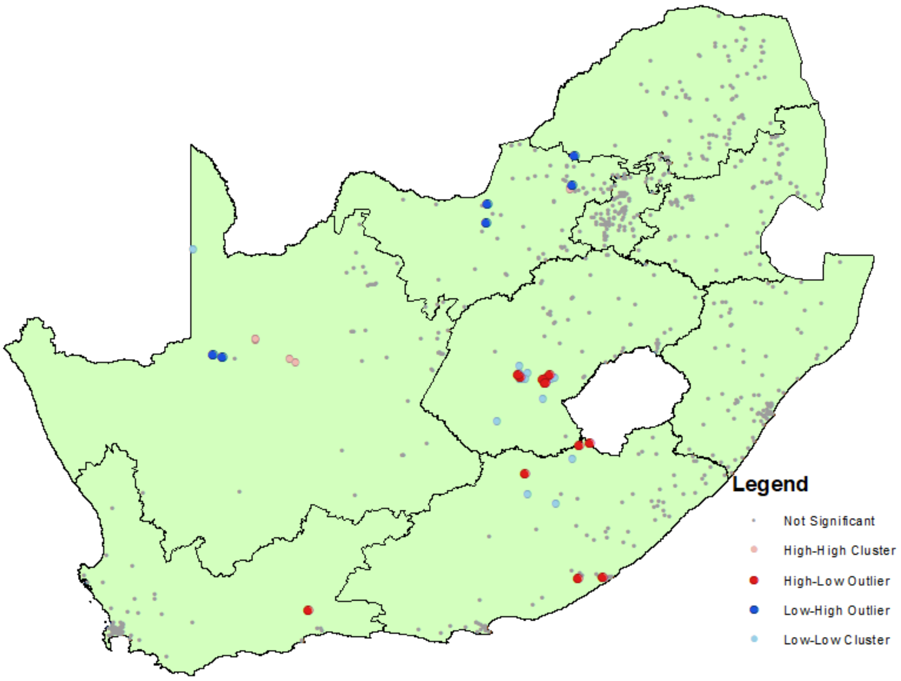

Figure 5.4: Anselin local Moran's I applied to the generalized additive models for non-diabetics

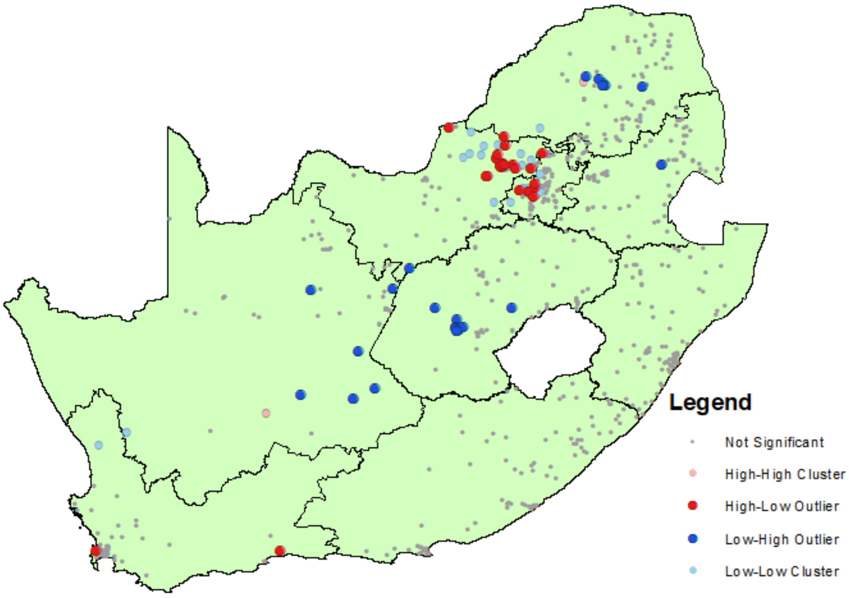

Figure 5.5: Anselin local Moran's I applied to the generalized additive model for diabetics

As spatial autocorrelation is no longer significant in the two separate models after the inclusion of the spline effects of longitude and latitude we therefore can conclude that the addition of those spline effects into a multinomial model would be sufficient to account for spatial autocorrelation in the residuals. The resulting model being a multinomial generalized additive mixed model. The final results for the multinomial generalized additive mixed model are given in Table 5.1. Laplace approximation was used for maximum likelihood estimation. Significant effects included age, highest education level, BMI, Rohrer's Index, waist circumference, waist-to-height ratio, haemoglobin level, taking high blood pressure medication, perception of health, approach towards salt consumption, consumption of fruit juice the previous day, the interaction of waist-to-height ratio and salt consumption, waist-to-height ratio and BMI, consumption of salt and fruit juice, perception of health and fruit juice consumption, highest education level and age as well as the non-linear effect of latitude.

Table 5.2 gives the estimated odds ratios and their 95% confidence intervals for variables not included in any interaction effects. Considering non-diabetics vs pre-diabetics: for a unit increase in Rohrer's index, individuals were less likely to be non-diabetic rather than pre-diabetic (OR=0.899; 95% CI:0.841-0.960). Similarly, for a unit increase in waist circumference, individuals were 0.959 times less likely to be non-diabetic rather than pre-diabetic (95% CI: 0.930-0.989). For a unit increase in haemoglobin level, individuals were 1.109 times more likely to be non-diabetic rather than pre-diabetic (95% CI: 1.051-1.172). Considering diabetics vs pre-diabetics: Black/Africans were more likely to be diabetic (OR=1.429; 95% CI: 1.032-1.978) compared to other race groups. Individuals taking high blood pressure medication were 1.444 times more likely to be diabetic than pre-diabetic (95% CI: 1.167-1.786).

Table 5.1: Type 3 results for multinomial generalized additive mixed model

Main Effects	**F value**	**P-value**
Longitude bspline	1.55	0.0990
Latitude bspline	4.02	< .0001*
Gender	2.78	0.0622
Race	2.64	0.0711
Age	25.72	< .0001*
EdLevel	4.40	0.0002*
BMI	4.53	0.0012*
Rohrers’ Index	6.16	0.0021*
Waist circumference	3.97	0.0189*
WtHR	3.30	0.0370*
Haemoglobin level adjusted for altitude and smoking	11.80	< .0001*
BP	1.15	0.3162
Taking high BP medication	6.11	0.0022*
Taking medication	2.17	0.1141
Perception of health	8.64	0.0002*
Consumption of processed foods	2.36	0.0944
Salt consumption approach	4.59	0.0102*
Consumption of fruit the previous day	0.44	0.6435
Consumption of vegetables the previous day	0.97	0.3802
Consumption of sugary drinks the previous day	0.68	0.5086
Consumption of fruit juice the previous day	3.90	0.0204*
Smoking the previous 24hrs	1.89	0.1510
Wealth index Z-score	1.62	0.1987
Interaction Effects		
WtHR & Salt consumption approach	7.65	0.0005*
WtHR & BMI	4.62	0.0010*
Salt consumption approach & Fruit juice consumption	7.06	0.0009*
Perception of health & Fruit juice consumption	8.03	0.0003*
Education level & Age	4.61	0.0001*

$*$ significant at 5% level of significance

Table 5.2: Multinomial generalized additive mixed model results

Variable	**Non-diabetic**	**Diabetic**
	OR (95% CI)	**OR (95% CI)**
Gender (ref=male)		
Female	1.368 (1.023-1.830)	0.913 (0.729-1.143)
Race (ref=other)		
Black/African	0.873 (0.557-1.370)	1.429 (1.032-1.978)
Rohrer′s Index	0.899 (0.841-0.960)	0.962 (0.925-1.001)
Waist circumference	0.959 (0.930-0.989)	1.006 (0.985-1.026)
Haemoglobin level adjusted for altitude and smoking	1.109 (1.051-1.172)	0.947 (0.909-0.986)
Blood pressure (ref=normal)		
Abnormal	0.907 (0.720-1.141)	1.098 (0.937-1.287)
Taking high blood pressure medication (ref=no)		
Yes	0.927 (0.650-1.323)	1.444 (1.167-1.786)
Taking Medication (ref=no)		
Yes	0.753 (0.540-1.051)	1.111 (0.908-1.360)
Household′s consumption of processed foods	0.909 (0.807-1.024)	0.929 (0.853-1.012)
Household′s consumption of fruit the previous day (ref=no)		
Yes	1.050 (0.855-1.290)	0.941 (0.805-1.012)
Household′s consumption of vegetables the previous day (ref=no)		
Yes	0.976 (0.797-1.195)	1.114 (0.951-1.305)
Household′s consumption of sugary drinks the previous day (ref=no)		
Yes	0.911 (0.736-1.127)	1.059 (0.899-1.248)
Smoking the previous 24hrs (ref=no)		
Yes	1.217 (0.929-1.593)	0.868 (0.680-1.108)
Wealth index factor score combined	1.037 (0.902-1.193)	1.099 (0.990-1.221)

Figures 5.6 to 5.9 display the effects of the interaction terms in the multinomial generalized additive mixed model. From Figure 5.6, it can be observed that an increase in waist-to-height leads to an increased chance in being non-diabetic rather than pre-diabetic. Approach towards salt consumption is noted as positive for individuals who have or believe they should decrease their salt consumption and negative otherwise. At a waist-to-height ratio of over 0.5, a negative approach towards salt consumption leads to a higher chance of being non-diabetic rather than pre-diabetic when compared to a positive approach towards salt consumption. Individuals with an increased waist-to-height ratio were less likely to be diabetic rather than pre-diabetic. Similar interactions can be seen in Figures 5.7 to 5.9.

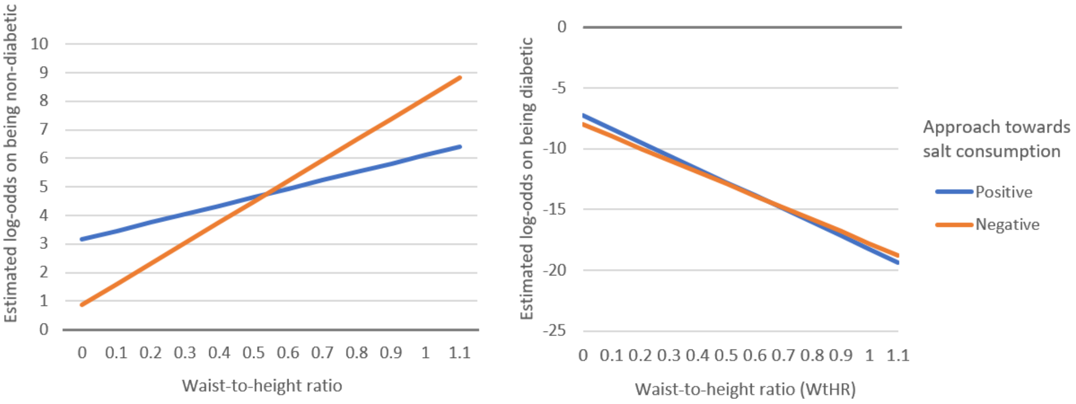

Figure 5.6: Interaction plot of waist-to-height ratio and salt consumption

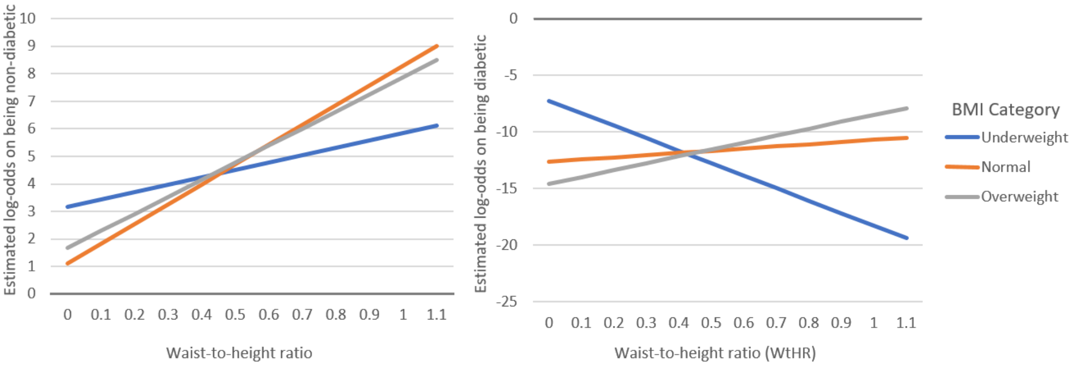

Figure 5.7: Interaction plot of waist-to-height ratio and BMI category

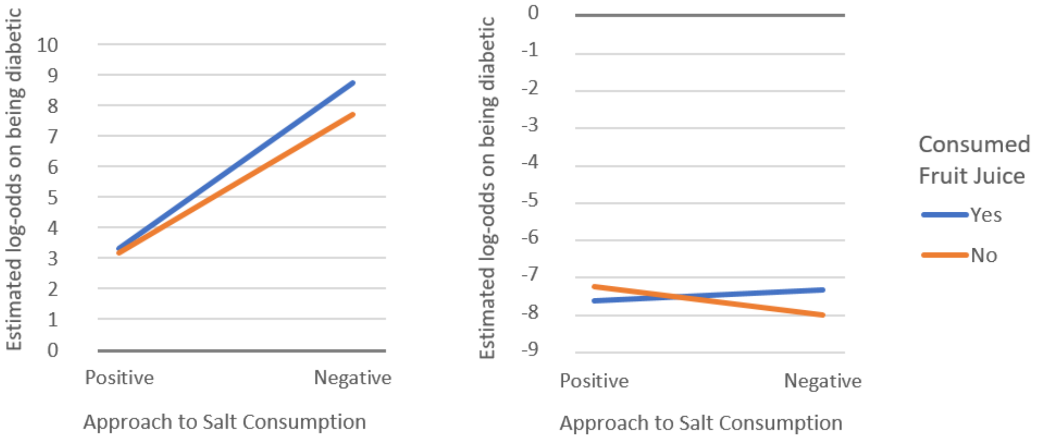

Figure 5.8: Interaction plot of approach towards salt consumption and consumption of fruit juice

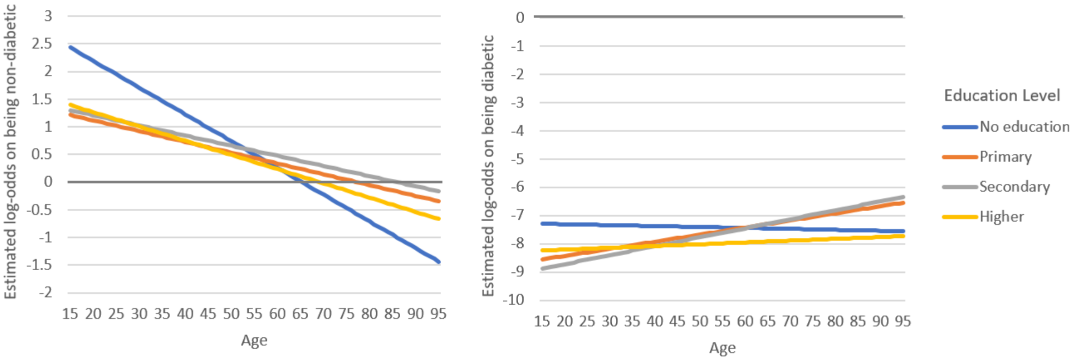

Figure 5.9: Interaction plot of age and education level

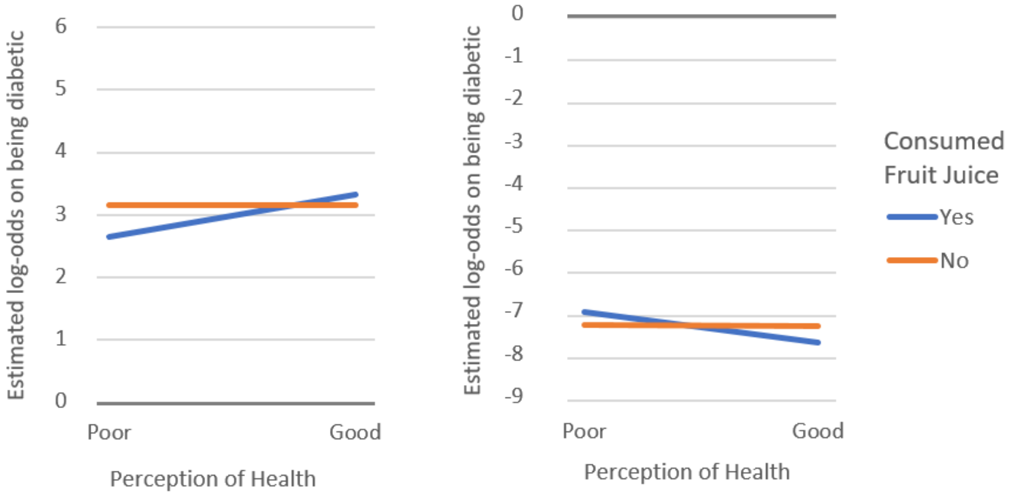

Figure 5.10: Interaction plot of perception of health and consumption of fruit juice

The predicted probabilities of diabetic status for each observation was extracted from the linear predictor. The diabetic status that yielded the highest probability was then the predicted outcome. The classification can be seen in Table 5.3 on the next page. This surface correlation adjusted model was found to be the best fit with an accuracy of 70.8%.

The last chapter in this thesis summarises the important findings, and discusses the accuracy of each statistical method considered for classifying a person's diabetic status. This chapter also discusses the limitations to the thesis as well as recommendations for future research.

Table 5.3: Confusion matrix for the multinomial generalized additive mixed model

Observed	Predicted			Total
	Non-diabetic	Pre-diabetic	Diabetic	
Non-diabetic	62	667	13	742
Pre-diabetic	24	4057	211	4292
Diabetic	3	966	439	1408
Total	89	5690	663	6442

Chapter 6

Discussion and Conclusion

This study aimed to investigate the prevalence and risk factors associated with diabetes in the South African population aged 15 years and older. Based on the data used in this study, the majority of individuals are facing, or are soon to face, the effects of this disease, where 22% were found to be diabetic and 67% found to be pre-diabetic. This high prevalence of pre-diabetes suggests that there will likely be a high prevalence of T2DM if no measures are put in place to prevent this. From the exploratory analysis, it was seen that both females and males had the peak diabetes prevalence in 70-100 year old participants. This result is similar to the study done by King et al. (1998) who reported that in developed countries, diabetes predominantly occurs in older age groups (65 years and older). The positive association of the wealth index score factor and being non-diabetic rather than pre-diabetic contradicts the study done by Peer et al. (2012) among urban-dwelling black South Africans. The significant association of BMI, Rohrer′s Index and waist circumference with diabetes concurs with findings from other studies, specifically where measures of central obesity were found to be more strongly associated with diabetes risk (Huxley et al., 2010; Motala et al., 2008; Peer et al., 2012).

T2DM has been found to be so closely related to obesity that the term *diabesity* has been coined by Fung (2018). Our results are consistent with the finding of Motala et al. (2008) where there is a positive association of waist circumference and diabetes.

Our results indicated that having smoked in the previous 24 hours was not a risk factor for being diabetic. However, smoking while diabetic is the strongest risk factor for peripheral vascular disease caused by atherosclerosis of the large blood vessels supplying the legs. Thus, the two together can accelerate the progression of peripheral vascular disease and ultimately could lead to the need for amputation (American Diabetes Association et al., 2003). Individuals taking high blood pressure medication were found to be at an increased risk for diabetes. However, according to Brunström & Carlberg (2016), such medications should be used with caution as diabetic individuals with a systolic blood pressure less than 140mm Hg that were on antihypertensive treatment were at an increased risk of cardiovascular death . It was observed that diabetes prevalence differed among provinces. This may be due to the lifestyle (e.g. diet, access to health facilities, education, culture etc.) accustomed in the provinces.

The final objective of the study was to explore various statistical methods of classifying a person's diabetic status. Accuracy was calculated for the overall model fit. Sensitivity, specificity and precision were calculated according to each outcome: non-diabetic, pre-diabetic or diabetic. This is summarized in Table 6.1. Although the decision tree and random forest may appear to be better in terms of sensitivity, specificity and precision, these two methods could not fully incorporate the survey design of the data (Zhang, 2014). Therefore, results from the decision tree and random forest are not reliable (UCLA, 2020). The Bayesian neural network yielded the highest accuracy (76.3%) however, it does not take the complex sampling design into account nor corrects for spatial autocorrelation in the residuals SAS Institute Inc. (2015). In addition, these classification techniques do not assist in determining the significant risk factors, only which risk factors are important (Livingston, 2005; Kenton, 2020). Thus, regression models were also considered.

Table 6.1: Classification statistics for the different models fit across the three classes: non-diabetic (N), pre-diabetic (P) and diabetic (D)

Model	Accuracy	Sensitivity (%)			Specificity (%)			Precision (%)		
	(%)	N	P	D	N	P	D	N	P	D
Decision tree	70.0	4.2	91.7	35.8	100	26.0	92.5	100	72.0	56.4
Random forest	68.2	0	97.0	16.3	100	11.2	97.2	0	68.9	62.0
Bayesian neural network	76.3	38.1	87.9	60.9	96.8	56.2	92.0	60.5	80.0	68.1
Ordinal survey logistic regression	68.1	0	96.4	18.0	100	12.3	96.6	0	68.7	59.9
Multinomial survey logistic regression	68.6	0	94.5	25.7	99.9	17.6	95.1	0	69.6	59.2
Multinomial generalized additive mixed model	70.8	54.7	94.5	31.2	99.5	24.0	95.6	69.7	71.3	66.2

The survey logistic model was able to account for the sampling design. However, the residuals were found to be spatially autocorrelated and therefore, results from the model were not reliable. The multinomial generalized additive mixed model, which accounts for spatial autocorrelation by adjusting for surface correlation and accounts for the survey design, best fits the data owing to its accuracy being the greatest. This is unsurprising as this was the only model that was able to incorporate the sampling design as well as account for spatial autocorrelation in the data (Zhang, 2014).

Some limitations of this study include a lack of variables in the SADHS 2016 data that have been found to be modifiable risk factors in previous studies such as cholesterol (lipids) level (Motala et al., 2008), carbohydrate and fat consumption (Noakes, 2013; Malhotra et al., 2015; Fung, 2018; Feinman et al., 2015) where dietary variables in this study were more sugar-based. Consumption of foods and drink and whether having smoked cigarettes were only recorded from the previous 24 hours. Also, variables related to diet were on a household level rather than an individual level. In addition to these limitations, this study was also based on data from a cross-sectional design. Therefore, a temporal relationship between diabetic status and the risk factors considered could not be established.

The future direction of this study includes

1. Assessment of more classification methods, such as support vector machines and deep learning.
2. Examining the change in the prevalence of diabetes and pre-diabetes in the South African population over time, by making use of additional data collected after the 2016 SADHS.
3. Investigating the joint effects of other diseases with diabetes, such as heart disease, hypertension, and stroke, among others.
4. Assessing diabetes prevalence and how it differs across the provinces of South Africa. Further, estimating disaggregated statistics at the district municipality or local municipality level using small area estimation.

References

(2012). *Simplifying the Analysis of Complex Survey Data Using the SAS® Survey Analysis Procedures*. Westat, Rockville, Maryland, USA.

(2017).
URL http://www.diabetesatlas.org/across-the-globe.html

Agresti, A. (2010). *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons.

Allison, P. (1999). *Logistic Regression: Using the SAS System: Theory and Application*. SAS Institute.

American Diabetes Association, et al. (2003). Peripheral arterial disease in people with diabetes. *Diabetes care*, *26*(12), 3333–3341.

Anthony, B. A. (2002). Performing logistic on survey data with the new surveylogistic procedure. [in Proceedings of the 27th Annual SAS Users Group International Conference (SUGI 27)].

Ashwell, M., & Gibson, S. (2016). Waist-to-height ratio as an indicator of 'early health risk': simpler and more predictive than using a 'matrix' based on bmi and waist circumference. *BMJ Open*, *6*(3).
URL https://bmjopen.bmj.com/content/6/3/e010159

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*, *66*(4), 398–407.

AVERT (2018). HIV and AIDS in SOuth Africa. https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa. [Online; accessed March 2019].

Basu, S., Yoffe, P., Hills, N., & Lustig, R. H. (2013). The relationship of sugar to population-level diabetes prevalence: An econometric analysis of repeated cross-

sectional data. *PLOS ONE*, *8*, 1–8. [Online; accessed March 2019].
URL `https://doi.org/10.1371/journal.pone.0057873`

Beleites, C., Salzer, R., & Sergo, V. (2013). Validation of soft classification models using partial class memberships: An extended concept of sensitivity & co. applied to grading of astrocytoma tissues. *Chemometrics and Intelligent Laboratory Systems*, *122*, 12–22.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, (pp. 1171–1178).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., & Cutler, A. (2003). Manual–setting up, using, and understanding random forests v4.0. `https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf`. [Online; accessed November 2019].

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Brunström, M., & Carlberg, B. (2016). Effect of antihypertensive treatment at different blood pressure levels in patients with diabetes mellitus: systematic review and meta-analyses. *BMJ*, *352*.
URL `https://www.bmj.com/content/352/bmj.i717`

Chen, C. (2000). Generalized additive mixed models. *Communications in Statistics - Theory and Methods*, *29*(5-6), 1257–1271.
URL `https://doi.org/10.1080/03610920008832543`

Cliff, A., & Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical analysis*, *4*(3), 267–284.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, *4*(5), 613–617.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3121.1992.tb00605.x`

De Wit, S., Sabin, C. A., Weber, R., Worm, S. W., Reiss, P., Cazanave, C., El-Sadr, W., Monforte, A. d., Fontas, E., & Law, M. G. (2008). Incidence and risk factors for new-onset diabetes in HIV-infected patients: the data collection on adverse events of anti-HIV drugs (d: A: D) study. *Diabetes care*, *31*(6), 1224–1229.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. 96. Sage.

Feinman, R. D., Pogozelski, W. K., Astrup, A., Bernstein, R. K., Fine, E. J., Westman, E. C., Accurso, A., Frassetto, L., Gower, B. A., McFarlane, S. I., et al. (2015). Dietary carbohydrate restriction as the first approach in diabetes management: critical review and evidence base. *Nutrition*, *31*(1), 1–13. [Online; accessed March 2019].

Fung, J. (2018). *The diabetes code*. Greystone Ltd.

Gill, J. (2000). *Generalized Linear Models: A Unified Approach*, vol. 134. SAGE Publications.

Green, A. (2017). Diabetes risk because of status. `https://www.news24.com/SouthAfrica/Local/City-Vision/diabetes-risk-because-of-status-20170503`. [Online; accessed March 2019].

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge University Press.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Chapman and Hall/CRC.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*, vol. 398. John Wiley & Sons, 3rd ed.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Huxley, R., Mendis, S., Zheleznyakov, E., Reddy, S., & Chan, J. (2010). Body mass index, waist circumference and waist: hip ratio as predictors of cardiovascular risk—a review of the literature. *European journal of clinical nutrition*, *64*(1), 16.

IDF (2019). What is diabetes. `https://www.idf.org/aboutdiabetes/what-is-diabetes.html`. [Online; accessed March 2019].

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Kenton, W. (2020). Statistical significance. [Online; accessed February 2020].
URL `https://www.investopedia.com/terms/s/statistically_significant.asp`

King, H., Aubert, R. E., & Herman, W. H. (1998). Global burden of diabetes, 1995–2025: Prevalence, numerical estimates, and projections. *Diabetes Care*, *21*(9), 1414–1431.
URL `https://care.diabetesjournals.org/content/21/9/1414`

Kumari, M., & Godara, S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction 1. [Online; accessed February 2020]. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.219.6038`

Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2), 1797–1801.

Levitt, N. S., Katzenellenbogen, J. M., Bradshaw, D., Hoffman, M. N., & Bonnici, F. (1993). The prevalence and identification of risk factors for niddm in urban africans in cape town, south africa. *Diabetes Care*, *16*(4), 601–607. [Online; accessed March 2019].
URL `http://care.diabetesjournals.org/content/16/4/601`

Liu, Y., Shi, W., & Czika, W. (2017). Building bayesian network classifiers using the hpbnet procedure. [Online; accessed October 2019].
URL `https://support.sas.com/resources/papers/proceedings17/SAS0474-2017.pdf`

Livingston, F. (2005). Implementation of breiman's random forest machine learning algorithm. *ECE591Q Machine Learning Journal Paper*, (pp. 1–13).

Loh, W.-Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, (pp. 361–386).

MacGregor, A., Bamber, S., & Silman, A. (1994). A comparison of the performance of different methods of disease classification for rheumatoid arthritis. results of an analysis from a nationwide twin study. *The Journal of rheumatology*, *21*(8), 1420–1426.

Mahlakoana, T. (2018). How antiretrovirals have cut the HIV/AIDS burden on SA's economy. [Online; accessed March 2019].
URL `https://www.businesslive.co.za/bd/national/health/2018-05-28-how-antiretrovirals-have-cut-the-hivaids-burden-on-sas-economy/`

Malhotra, A., Noakes, T., & Phinney, S. (2015). It is time to bust the myth of physical inactivity and obesity: you cannot outrun a bad diet. *British Journal of Sports Medicine*, *49*(15), 967–968. [Online; accessed March 2019].
URL `https://bjsm.bmj.com/content/49/15/967`

McCulloch, C. E., Searle, S. R., et al. (2001). *Generalized, Linear and Mixed Models.*. Wiley, New York.

Meintjes, G., Moorhouse, M. A., Carmona, S., Davies, N., Dlamini, S., Van Vuuren, C., Manzini, T., Mathe, M., Moosa, Y., & Nash, J. (2017). Adult antiretroviral therapy guidelines 2017. *Southern African journal of HIV medicine*, *18*(1).

Millington, A. (2019). South Africa has just been ranked the unhealthiest country on earth. `https://www.businessinsider.co.za/most-unhealthy-countries-in-the-world-ranked-2019-3`. [Online; accessed March 2019].

Motala, A. A., Esterhuizen, T., Gouws, E., Pirie, F. J., & Omar, M. A. (2008). Diabetes and other disorders of glycemia in a rural South African community. *Diabetes Care*, *31*(9). [Online; accessed March 2019].

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Noakes, T. (2013). Low-carbohydrate and high-fat intake can manage obesity and associated conditions: Occasional survey. *South African Medical Journal*, *103*(11). [Online; accessed March 2019].

Noakes, T., & Sboros, M. (2017). *Lore of nutrition: challenging conventional dietary beliefs*. Penguin books.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175.

Peer, N., Steyn, K., Lombard, C., Lambert, E. V., Vythilingum, B., & Levitt, N. S. (2012). Rising diabetes prevalence among urban-dwelling black south africans. *PloS one*, *7*(9), e43336.

Peterson, B., & Harrell Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *39*(2), 205–217.

Pijoos, I. (2018). Diabetes claims around 2500 leg amputations a year in KZNl. `https://www.timeslive.co.za/news/south-africa/2018-11-11-diabetes-claims-around-2500-leg-amputations-a-year-in-kzn`. [Online; accessed March 2019].

Ramani, R. G., & Sivagami, G. (2011). Parkinson disease classification using data mining algorithms. *International journal of computer applications*, *32*(9), 17–22.

Reinehr, T. (2013). Type 2 diabetes mellitus in children and adolescents. *World journal of diabetes*, *4*(6), 270. [Online; accessed March 2019].

Sarle, W. S. (1994). Neural networks and statistical models.

SAS Institute Inc. (2015). SAS Enterprise Miner 14.1: High Performance Procedures. `https://documentation.sas.com/?cdcId=emlearncdc&cdcVersion=1.0&docsetId=emex&docsetTarget=titlepage.htm&locale=en`. [Online; accessed June 2019].

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*, vol. 391. John Wiley & Sons.

STATS SA (2016). Mortality and causes of death in South Africa, 2016: Findings from death notification. `http://www.statssa.gov.za/publications/P03093/P030932016.pdf`. [Online; accessed March 2019].

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 307.

The Lancet (2017). Burden of diabetes set to increase across sub-Saharan Africa, potentially diminishing health gains of recent years. `https://www.sciencedaily.com/releases/2017/07/170706072639.htm`. [Online; accessed March 2019].

UCLA (2020). Introduction to survey data analysis. `https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consult` [Online; accessed February 2020].

WHO (2018). Diabetes. `https://www.who.int/news-room/fact-sheets/detail/diabetes`. [Online; accessed March 2019].

Williams, R. (2016). Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology*, *40*(1), 7–20.

Zhang, A. (2014). SASware Ballot Ideas: A Weight statement for decision tree node SAS EM. `https://communities.sas.com/t5/SASware-Ballot-Ideas/A-Weight-statement-for-decision-tree-node-SAS-EM/idi-p/220134`. [Online; accessed February 2020].