National College of Ireland

Machine Learning algorithms in Health Questionnaires: Multiple Correspondence Analysis and Classification model

MSc Research Project
Data Analytics

Louise Blake
X13110535

School of Computing
National College of Ireland

Supervisor: Barry Haycock

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Louise Blake
Student ID:	X13110535
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Barry Haycock
Submission Due Date:	11/12/2017
Project Title:	Machine Learning algorithms in Health Questionnaires: Multiple Correspondence Analysis and Classification model
Word Count:	9550

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Machine Learning algorithms in Health Questionnaires: Multiple Correspondence Analysis and Classification model

Louise Blake
X13110535
MSc Research Project in Data Analytics

10th December 2017

Abstract

This paper describes the development of a unique multiple disease classification tool to expose and pre-screen for chronic diseases and can be applied to individual surveys. Through data mining of NHANES questionnaires, a transparent and straightforward model is developed which be could apply to future survey data. We show machine learning and dimensional reduction can be beneficial to survey data to determine the risk of multiple chronic diseases in individuals surveyed. The results are used to create a prototype tool that can predict the presence of multiple chronic diseases through fundamental questions. The number of questions are reduced to 5 questions to achieve an acceptable result. The researchers know of no tool currently available that delivers this kind of functionality.

1 Introduction

> "Health is a resource for everyday life, not the object of living. It is a positive concept emphasizing social and personal resources as well as physical capabilities"(World Health Organization and others; 1986).

When applied to data classification, machine learning algorithms are valuable tools in the discovery of features and classes. Classification algorithms apply to many domains for example medical diagnosis and credit card fraud. The objective of this research is to utilise machine learning algorithms to predict a classification of a person at risk of a single or multiple chronic diseases using questionnaire survey data. The key deliverable is a model to predict a classifier and to evaluate the performance of the model for the classification. The multi-classifier in this research comprises of four chronic diseases: arthritis, diabetics, hypertension and cholesterol conditions. Another deliverable is the creation of an application with a user interface to ask the questions and predict the classifier in the form of a Multi-Chronic Disease Self-Assessment Tool (MCDSAT).

The motivation for this research is to gain knowledge on health risks and non-communicable diseases (NCDs) from lifestyle risk factors. The four chronic diseases arthritis, diabetics, hypertension and cholesterol, have known associations with lifestyle

risk factors. These diseases are also a leading cause of health loss and mortality. Annually 40 million people die from NCDs; this is globally equivalent to 70% of all deaths (Forouzanfar et al.; 2016). Other connected terms are lifestyle and wellness used in the context of the papers reviewed. Examples of risk factors that increase an individuals' likelihood of developing a disease include behaviour such as diet, smoking, physical activity, and alcohol all of these are modifiable risk factors. Risk factors that, cannot include age, race or family history (World Health Organization; 2015). Surveys are one of the mechanisms used to discover relationships between health and lifestyle, as an informing tool to detect a persons risk assessment.

The scope of this research is limited, as it does not include all the many classification algorithms available. The questionnaire data excludes laboratory test results and family history data. The final developed MCDSAT application is not sufficient to provide a diagnosis or prognostic model but shows that one is achievable.

This paper brings together literature reviewed. Chapter 1 introduces risk factors, non-communicable diseases and the motivation for the research. Chapter 2 presents the literature reviewed; the machine learning approaches taken for reduction of questionnaire data, approaches to feature selection and classification and online self-assessment tools. Chapter 3 is the methodology applied while chapter 4 is multiple correspondence analysis and section 5 presents the evaluation, results and discussion of the classification model. Chapter 6 is the implementation and deployment and finally chapter 7, the conclusions and future work.

2 Literature review

In this chapter, the related work is reviewed. This review identifies classification algorithms, classification for prediction and dimensional reduction and online self-assessment tools that are used for predicting health issues.

2.1 Classification algorithms

Domains benefit from using classification algorithms such as medical diagnostics, epidemiology and health studies where the reliability of the diagnosis is critical. Different types of classification algorithms; logistic regression, support vector machines, decision trees and k-nearest neighbour classifier each have their own merits. For example, regression tree models can apply when the dependent variable (predicted variable) is numeric, classification when the dependent variable is categorical. The objective of the research is to predict categorical class label on a multi-classifier. Classification algorithms that perform on categorical data include; decision tree classifiers, rule-based classifiers (Apriori algorithms), Neural Networks and Naïve Bayes classifiers. The choice of which classification method to use is a consideration.

Heikes et al. (2008) developed a screening tool for diabetes and pre-diabetes and applied a two algorithm strategy - logistic regression for probability and classification strategy to build a decision tree. CART algorithm is a popular tree-based classifier used for classification and regression analysis. The CART algorithm is used for the decision tree while the data characteristics are delivered from a demographic and laboratory analysis dataset. It is one of the algorithms employed for a response variable that is categorical in its characteristics. CART was developed by Breiman et al. (1985) in the early 1980'S. Sathyadevi (2011) uses CART to classify hepatitis disease diagnosis. The advantages of

CART are that it accommodates missing values in the datasets, and it employs cross-validation (cv) to assist in sizing the tree. CART is a tree-based modelling technique, and a benefit of a tree classification model is it interpretability, it is also non-parametric. As an algorithm CART uses recursive partitioning, it grows the tree and then prunes the tree (Loh; 2011). It also produces rules that are easy and logical to explain; this is an advantage compared with Neural Network models.

2.2 Classification for prediction

This section presents the literature reviewed around classification algorithms for prediction and the evaluation measures in this domain. A study of classification techniques by Loh (2011) compares the features of six classification methods, C.45, CART, CHAID, CRUISE, GUIDE and QUEST. The feature comparison includes, missing values, node models, variable ranking, pruning and split types and CART algorithm have a presence indicated for each of these features. Loh (2011) suggests that compared to GUIDE, CART has less accuracy and computational speed as C4.5. Alizadehsani et al. (2013) compares classification algorithms Bagging SMO, Naïve Bayes, SMO, and Neural Network for performance results to establish and identify the most useful features of coronary artery disease. The performance measures produce a selection of features, and the four classification algorithms use the matrices; accuracy, sensitivity and specificity, confusion matrices and ROC curves (Alizadehsani et al.; 2013). Alizadehsani et al. (2013) reports that better results to achieve predicting coronary artery disease, for the performance matrices accuracy, sensitivity and specificity of algorithms examined when using feature selection and not all of the features. Liu et al. (2014) reports on the benefits of algorithms examined they can extract decision rules and select important features. They examine the evaluation of algorithms; Support Vector Regression and random forests comparing the metrics accuracy, interpretation, and robustness. An algorithm can perform well but does not interpret metrics rules and variable numbers well (Liu et al.; 2014).

Kumari et al. (2014) applies BN to predict diabetes with the objective to class a person as diabetic, non-diabetic or pre-diabetic. (Kumari et al.; 2014) reports an accuracy of approximately 99% using the BN for a multiclass predictive model. Chaurasia and Pal (2013) applies some classifier techniques, Naïve classifiers, bagging and J48 decision tree in the diagnosis of heart diseases and the risk factors associated with heart disease. The results indicate that bagging produces slightly more accuracy performance than the other two classifiers. Chaurasia and Pal (2013), Alizadehsani et al. (2013) and Polat and Güneş (2007) use different classifiers and achieved varying results for the measurement of accuracy, sensitivity, and specificity. Loh (2011), Alizadehsani et al. (2013), Kumari et al. (2014), Polat and Güneş (2007), and Chaurasia and Pal (2013) compared prediction performance of their algorithms including Support Vector Machines, Artificial Neural Network, CART, C.45, Multilayer Perceptron, Bayesian Networks and Random Forests are some of these. For this research, a set of rules is a requirement, and human interpreted, this is taken into account when choosing an algorithm. The measurement of interpretation is by variables and rules produced in a decision tree. A decision tree algorithm such as CART produces a single tree that when pruned, produces rules that can interpret and provide predictive accuracy.

2.3 Classification for feature selection

In this section literature covering classification algorithms is reviewed for dimensional reduction. The selection of features can cause a problem when building a multiclass model, and some classification models have the variable importance feature which is a benefit when data requires dimensional reduction. Polat and Güneş (2007) applied the C4.5 decision tree algorithm to reduce the dimensions of a heart and hepatitis disease dataset, then employed a pre-processing fuzzy weighted method and artificial immune recognition system (AIRS). The accuracy level achieved by was 92.59% however they found that the AIRS classifier does not manage the multiclass problem very well, as it does not classify all the points but includes other methods such as Kernel functions. Alizadehsani et al. (2013) and Polat and Güneş (2007) apply different approaches for the selection of important features. The approaches include a hybrid model with both domain expert and ML by Alizadehsani et al. (2013) and ML, the C4.5 Polat and Güneş (2007). Li et al. (2004) approach to tissue classication based on gene expression is to build a multiclass classifier by first selecting the features the classication method. A problem noted by Li et al. (2004) is high dimensional of the data, and if the classication is a multiclass accuracy appears to degenerate with increasing classes, this is a consideration for the research as the data has high dimensionality.

2.4 Dimensional reduction

For dimensional reduction, options available include Principal Component Analysis (PCA). The implementation of PCA requires the questionnaire categorical (factor levels) converted into binary dummy variables (one hot encoding). An alternative option for conversion of categorical variables into numbers includes deviation, Helmert, orthogonal polynomial (UCLA:Statistical Consulting Group; 2011). Sourial et al. (2010) discusses the benefit of applying both PCA and factor analysis in epidemiological studies and noted that these techniques are designed for use with continuous variables. Another technique is Multiple Correspondence Analysis, which is part of a family of methods developed for Correspondence Analysis. MCA is a relevant methodological approach for exploring the individual response categories of the categorical variables, and the functionality includes, descriptive data analytic techniques for multivariate, which makes it a good candidate for this research and the dimensional reduction stage of the process.

Thus MCA is a suitable technique for qualitative data while PCA is suitable for quantitative data. Sourial et al. (2010) and Costa et al. (2013) reports on studies in epidemiology, health and medical social studies where both qualitative and quantitative data are present and benefit from MCA. The difference between PCA and MCA is the way PCA treats the column relations, by decomposing their covariance matrix and treating the rows as cases. In MCA the columns and row are treated at the same time, hence the variables and categories results are accessible to interpret for the coordinates as there are more details available (Costa et al.; 2013),(Josse and Husson; 2012). Costa et al. (2013) applies both MCA and PCA for reducing and exploring data from cognitive, clinical, physical, and lifestyle variables and to investigate relationships to ageing. They apply PCA to reduce the information for neurocognitive data, this data comprises of cognitive test results, and the MCA technique is used to explore the data.

The reported benefits of MCA include a summary of analysis and visualisation of the dataset produced for variables with multiple categories (Costa et al.; 2013). The questionnaire data is predominantly categorical variables with continuous variables, trans-

formed into categorical using binning techniques. MCA can accommodate the analysis of categorical variables and it can also treat missing values as an added level, categories which are sparse examine or treat it as another category. MCA is part of the pre-processing stage of the analysis (Kassambara and Mundt; 2016). The MCA technique provides analysis of categorical data, with results consisting of a set of eigenvalues, column and row coordinates and the Cos2 (also known as squared correlations), and the quality of representation of a variable category or an individual in n dimensions. The contribution values are output to process post-hoc and sorted to obtain the set of features contributing the most and those features contributing the least. Another reported benefit of MCA is that it can visualise categories and individuals using graphical plots (Sourial et al.; 2010), (Kassambara and Mundt; 2016) (Lê et al.; 2008) (Husson, Lê and Pagès; 2017).

2.5 Online self-assessment tools for predicting health issue

In this section online self-assessment tools (SAT) are reviewed. The purpose of reviewing online SAT is to discover if data mining methodologies are applied. A vast array of health self assessment tools is available on line. The foundations of the examples in Table 1 are domain expert research, data accumulated over a number of years and focus on single medical conditions (Chaurasia and Pal; 2013), (American Diabetes Association and others; 2004), (Gellish et al.; 2007), (National Heart, Lung, and Blood Institute and others; 2013).

Table 1: Online Self Assessment Tools (SAT)

Tool.Name Online Publisher	Literature Author(s),Published	Chronic Disease	Comment
Cardiovascular Lifestyle Calculator Risk Score Harvard T.H. Chan School of Public Health (2017)	Chaurasia and Pal (2013)	Cardiovascular health	A quiz to evaluate how your current lifestyle habits. Information collected, 5 areas; smoking, weight, physical activity, alcohol use, and diet.
Type2_ Diabetes Risk Test American Diabetes Association (2017)	American Diabetes Association and others (2004)	Type 2 Diabetes	The online tool is an adaptation of a paper survey from the American Diabetes Associations Diabetes Risk Test. It is years of domain expert research, and has a scoring methodology for various factors such as weight and age.
Target Heart Rate Calculator National Heart Foundation of Australia (2017)	Gellish et al. (2007)	Heart Rate	"Target Heart Rate (THR) range values are often calculated to ensure exercise intensity is maintained at a desired level. This calculator automatically calculates THR ranges" (National Heart Foundation of Australia; 2017).
Risk Assessment Calculator American Council on Exercise (2017)	National Heart, Lung, and Blood Institute and others (2013)	Heart attack risk	This calculator uses the Framingham risk score to predict your chance of having a heart attack within the next 10 years.

3 Methodology

This section describes the methodology applied for the data mining process and the Machine learning algorithms.

3.1 CRISP-DM

The methodology chosen for this research is Cross-industry standard process for data mining (CRISP-DM). It has six phases and provides a framework to set a hypothesis and objectives to analyse the data (Figure 1). The advantages of CRISP-DM methodology are flexibility and the iterative recurring process in the methodology (Azevedo and Santos; 2008).

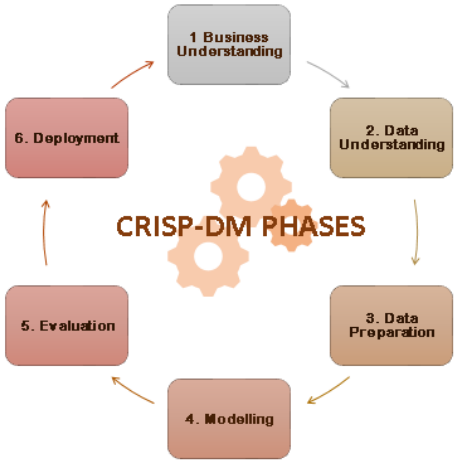

Figure 1: The phases of CRISP-DM (Azevedo and Santos; 2008)

3.2 Business Understanding

3.2.1 Data understanding

The data files are sourced from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 survey period. The NHANES surveys are collected on a two-year cycle and include demographic, laboratory, health questionnaire and physical examinations components. The components of the research include demographic and health questionnaire used to build a Master dataset. The Master dataset has resulted from 3707 participants. The files are merged into a single dataset in an MS SQL database. The selection of questions (variables) extracted from the components, Demographics and Questionnaire are listed in Table 2. The dataset characteristics are multivariate categorical responses, discrete data with a few transformed continuous variables. A complete listing of variables sourced from the National Health and Nutrition Examination Survey Data Documentation, Codebook, and Frequencies (Centers for Disease Control and Prevention; 2017).

3.2.2 Data challenge

The NHANE datasets have an extensive set of questionnaires, hence the need for dimensional reduction. Feature selection and extraction are important stages in the machine learning process and a few methods are used to assist in deciding the variables that are important and which are not. The first challenge in the data selection is to decide which four chronic diseases to target, the second challenge is to determine the most important variables to predict the target classifier, which is based on four chronic diseases. These decisions occur at the pre-processing stage.

Table 2: NHANES Demographic and Questionnaire components

Data File Name	NHANES 2013-2014 Data File, Published	NHANES 2009-2011 Data File, Published
Demographic	DEMO H, Oct. 2015	DEMO F, Sept. 2011
Alcohol Use	ALQ H, Mar. 2016	ALQ F, Jan. 2012
Blood Pressure Cholesterol	BPQ H, Oct. 2015	BPQ F, Sept. 2011
Diabetic	DIQ H, Oct. 2015	DIQ F, Sept. 2011
Medical Conditions	MCQ H, Oct. 2015	MCQ F, Sept. 2011
Osteoporosis	OSQ H, Oct. 2015	OSQ F, Jan. 2012
Smoking Cigarette Use	SMQ H, Revised Sept. 2016	SMQ F,Updated May, 2015

3.3 Framework for data processing

The data preparation has some stages and Figure 2 illustrates the framework for the pre-processing pipeline stages. Data files are extracted from the source and loaded into staging tables in a Microsoft SQL Database. The data is checked for inconsistencies such as duplicate indices. The index variable name is SEQN_ID; this is the Respondent sequence number and available for each dataset observations. This variable is used to create a merged database called the Master Questionnaire dataset. Each file in Table 2 from the NHANES 2013-2014 has many variables hence subsets of variables are used in the Master Questionnaire dataset. Post merging the SEQN_ID is dropped from the dataset as it has no value for the machine learning models.

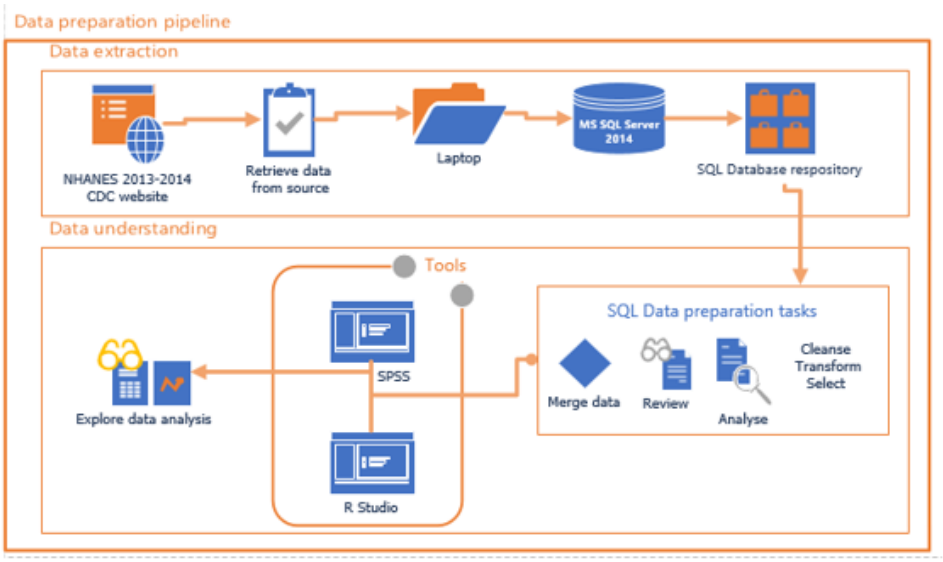

Figure 2: Framework for data processing pipeline

3.3.1 Data pre-processing

Before applying the data mining model the data requires pre-processing. This stage is essential to inspect the data characteristics as the content dictates the required feature selection and data transformations. The pre-processing stage includes transformations, encoding of values, imputation, features engineering, binning, and normalisation. The Centre for Disease Control releases the raw data source file in the SAS transport file (*.xpt) format, a numerically coded file. A sample of a variable from the components is provided Table 3.

Table 3: NHANES Sample variable characteristics pre-processing

Data File	Variable	Label	Value or Code
Alcohol	ALQ101	Had at least 12 alcohol drinks per year	1 = Yes
			2 = No
			0 = Missing
			7 = Refused
			9 = Dont Know

3.3.2 Recode sparse categories

To reduce the sparsity levels some of the levels in that categorical variable is re-coded. An example is where the Question variable has responses which include; Yes, No, Refused, Dont know and Missing responses. The variables are cleaned to get rid of the non-response categories by combining Refused, Dont know and Missing values. In Table 3, ALQ101 is converted from value 1 to Yes and 2 to No, all other answers NA.

3.3.3 Transformation

To convert numerical variables, a process of grouping into bins is required, hence a series of ranges are created for the conversion of continuous features. The continuous variables once binned are then removed/discarded. The same approach is applied to each of the age-related variables using bins of equal range approach, for example age binning ranges are 0, 1-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80 and over.

3.4 Exploratory Analysis

In SPSS data exploration of the Master dataset (seventy questions) includes visualisation of graphical plots forming histograms, boxplots and Q-Q plots. Statistical tests comprise of skewness values, and the two tests for normality Kolmogorov-Smirnov and Shipro-Wilkes. The observed p-values were $p< 0.000$ and the test for independence. Contingency tables and Chi-square tests for probabilities are used to evaluate whether there is a significant dependence between row and column categories p-value$<$ 2e-16. The observed results for the variables are statistically significantly associated (p-value).

3.5 Feature selection and extraction

The next stage in the process is to reduce the variables and the choice of chronic diseases. Employed methods include features selection and feature extraction. The initial criteria for a question are that the data is from self-reporting questions. The NHANES components comprised of self-report survey questions, which are where the respondents read the question and select a response. Some of the questionnaire components relate to examination and laboratory results, which are out of scope. Choosing questions from the self-reporting questions is a manual process and is dependent on the dataset. It is not a bias-free process as the selected data is conditional and excludes laboratory and examination questions. The initial set of questions totalled seventy variables which included twenty chronic diseases.

3.5.1 Feature selection

Reducing the dataset further is the next stage. Features are dropped using a threshold based on percentages of missing values $>=90\%$ and the sparse proportion of yes answers

<=5%. The structure of the dataset variables is fundamental, and as a high percentage >=90% of missing data values could impact the quality of imputation. A design decision is to limit imputation on the dataset to avoid over fitting issues. Also, the sparse levels are consolidated and recorded (Conway and Huffcutt; 2003). Illustrated in Figure 3 is the framework for the feature selection and extraction. A summary of the discarded and

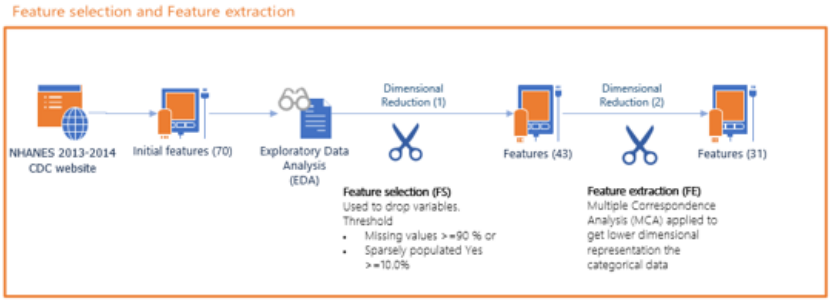

Figure 3: Feature selection and extraction pipeline

the retained chronic diseases is presented in Table 4. Statistical analysis using frequency and MCA techniques assist in establishing the final set of chronic diseases to retain and which question variables to drop. The top four chronic diseases were selected to create a composite target variable. The composite target variable comprises of the four target chronic diseases selected are Arthritis (A), Cholesterol (C), Diabetes or Diabetes (Borderline) (D), and Hypertension BP (H).

Table 4: Twenty chronic medical conditions, frequency and missing values percentages.

Code	Description	Yes %	MV %	Dropped variable
BPQ020	Ever told had Hypertension BP [1]	49.2	0.1	
BPQ080	Ever told blood cholesterol level high [1]	46.4	0.8	
MCQ160A	Ever told you had arthritis [1]	36.4	0.3	
DIQ010	Ever told have Diabetes or sugar diabetes [1]	18	0.1	
MCQ010	Ever told had asthma	14.2	0.1	
MCQ160M	Ever told you had thyroid problem	13.4	0.2	
OSQ060	Ever told had osteoporosis or brittle bones	8.5	0.3	
MCQ160K	Ever told you had chronic bronchitis	6.6	0.2	
MCQ160C	Ever told had coronary heart disease	5.9	0.4	X
MCQ160E	Ever told you had a heart attack	5.8	0.1	X
MCQ160N	Ever told you that you had gout	5.8	0.1	X
DIQ160	Ever told you have prediabetes	5.5	22.1	X
MCQ160L	Ever told you had any liver condition.	5.3	0.2	X
MCQ160F	Ever told you had a stroke	5.1	0.1	X
MCQ160b	Ever told had congestive heart failure	4.6	0.2	X
KIQ022	Ever told have weak or failing kidney	4.2	0.2	X
MCQ160D	Ever told you had angina or angina pectoris	3.4	0.1	X
MCQ160G	Ever told you had emphysema	2.3	0	X
MCQ082	Ever told that celiac disease or sprue	0.6	0.1	X

[1] One of the four target diseases.

3.5.2 Feature extraction

Feature extraction is a methodology used to transform the data from a high-dimensional set of variables to fewer dimensions using methods such as Multiple Correspondence Analysis.

3.5.3 Multiple Correspondence Analysis

A challenge in the dataset is that there were many variables to select. The purpose of dimensional reduction techniques is to reduce the dataset. Two aims of dimensional reduction are to compress the data into as few dimensions as possible and to decide which chronic diseases to predict and to produce the composite variable. The techniques applied to the dataset include statistical analysis and MCA. Subsequent to exploratory data analysis, the dataset is reduces from seventy to forty-five variables with 123 variable categories. Subsequent to MCA, the dataset is further reduced Table 5 presents the variables for the classification model. The output from MCA includes eigenvalues, variances, visualisations and the results for individuals and variables. The principle components explain which data accounts for most of the variation of original data and the highest possible variance. Another output is the measurement of distance using the cosine similarity (cos2). These measurements assist in the compression of the data into a few dimensions. The aim is to preserve 98% of the variance when applying dimensional reduction. The MCA is applied using the Factoextra package in R studio (Kassambara and Mundt; 2016). Section 4 discusses the results from MCA in more detail and further details, are available in the configuration manual.

Table 5: Twenty seven questionnaire variables

Cycle 2013-2014	Cycle 2009-2010	Recoded with partial description
RIDAGEYR	RIDAGEYR	d_age
RIAGENDR	RIAGENDR	d_gender
ALQ101	ALQ101	B_ALQ101.drink.12.minperyr
ALQ151	ALQ150 [2]	B_ALQ151.drink.4.or.more.daily
SMQ020	SMD020	B_SMQ020.smoked.100.in.a.lifetime
SMQ040	SMQ040	B_SMQ040.now.smoke.cigarettes
SMQ030	SMD030	B_SMQ030.age.started.smoking
MCQ365B	BPQ090C [2]	B_MCQ365B.increase.physical.activity
MCQ365C	BPQ090B [2]	B_MCQ365C.told.reduce.sodium.or.salt
MCQ365D	BPQ090A [2]	B_MCQ365D.told.reduce.calories
MCQ080	MCQ080	B_MCQ080.told.overweight
BPQ090D	BPQ090D	M_BPQ090D.prescription.cholesterol
BPQ100D	BPQ100D	M_BPQ100D.now.prescribed.lower.cholesterol
BPQ050A	BPQ050A	M_BPQ050A.now.taking.prescribed.HBP
BPQ040A	BPQ040A	M_BPQ040A.prescribed.BP.hypertension
BPQ090D	BPQ090D	M_BPQ090D.prescribed.cholesterol
DIQ050	DIQ050	M_DIQ050.taking.insulin.now
DIQ070	DIQ070	M_DIQ070.diabetic.pills.lower.blood.sugar
BPQ030	BPQ030	C_BPQ030.hypertension.BP.more.than.once
MCQ010	MCQ010	C_MCQ010.told.had.asthma
MCQ025	MCQ025	C_MCQ025.age.asthma
MCQ035	MCQ035	C_MCQ035.still.have.asthma
MCQ160K	MCQ160K	C_MCQ160K.chronic.bronchitis
MCQ160M	MCQ160M	C_MCQ160M.thyroid.problem
MCQ180M	MCQ170M	C_MCQ180M.Aae.thyroid.problem
MCQ170M	MCQ170M	C_MCQ170M.still.have.thyroid.problem
OSQ060	OSQ060	C_OSQ060.Osteoporosis

[2]Coding differences between Cycle 2009-2010 and Cycle 2013-2014.

3.6 Classification model

The data characteristics influence the choice of classification algorithm and the objective of the model at deployment stage. The dataset has a predictor with sixteen categories; it is a multivariate classifier comprising of four chronic diseases. At deployment stage the decision tree results are used to create rules.

Visualising the rules of the decision tree model can interpret easier (Breiman et al.; 1985). Decision trees can perform variable selection and can reduce data preparation by managing missing values and outliers (Loh; 2011). Decision trees do not require any assumptions of linearity in the data. A weakness of decision trees is that they can become complicated. The outcomes-based on expectations, which tree-based models do not fit as well for continuous variables predictors. Over-pruning is another issue; an aim is to minimise overfitting the tree by applying the cross-validation technique to the model at the building stage (Loh; 2011). Decision trees algorithms construct the tree; in a top-down manner, they are recursive and have a greedy characteristic.

The CART algorithm considers all possible subsets of variable categories which are good for high dimensional predictors and uses a binary split. How well the two classes split can be measured by construction of impurity, which is the degree of heterogeneity of the leaf nodes. The measurement of impurity is quantified by the following; Entropy, Gini Index and Classification Error. Impurity is a measure of Entropy and available in C4.5, and the Gini a measure in CART (Breiman et al.; 1985). Results are discussed in Section 5.

3.6.1 Test and train dataset

The data is split into test and train datasets using a function which inputs the data frame and divides it into two data frames one is named as train set, the other test set. A partition of the dataset using the random simple sampling technique split the dataset into subsets. A common method is to split the data ratio of 0.7, 70% training and 30% testing data, built under R version 3.4.2 (Tuszynski; 2014).

3.7 Model evaluation

In the literature reviewed in Section 2 measures for evaluation for classification models are referred to. This section discusses in more detail the purpose of evaluation and the various measurements that can be used to quantify the classification model.

The purpose of the evaluation stage is to determine if the model built is a good representation of the truth, using the test dataset to determine the performance of the algorithms. There are two evaluation attributes for classification models; accuracy is a measure of how often the model gets its predictions right and reliability is a measure of how consistent the model is with different data sets. The model is built using the train data set, and the tuned model applies to the test data. Three commonly used evaluation classification metrics are Precision, Recall, and F-measurement. In this research, other method differentiation is included such as Kappa. The Area Under ROC Curve is not applied as the ROC metrics are only suitable for two-class classification problems. The results for evaluation of the model classification are in Section 5 and further details are available in the configuration manual. The classification model CART evaluates for performance with comparison metrics from the random forest, and CART models produce the decision tree rules.

3.7.1 Cross-validation and Pruning

The model is fine-tuned to optimise the parameters and to increase its performance. The cross-validation technique is used to validate models such as classifiers. The method estimates how accurate the model performs with unseen data. The Complexity Parameter

(cp) of the tree and cross-validated error (x-error) are used to evaluate the decision tree (Therneau et al.; 2015). A built-in cross-validation function is available in the R package Rpart and can calculate the cross-validation error (x-error), the alternative is to calculate the output. The performance measures; root node error, rel_error, xerror column, can use to compute the optimum complexity parameter. The output from the three columns is used to calculate the cp to determine where to prune the tree (Therneau et al.; 2015). In this research, the cross-validation uses a repeated k-fold to produce the results in Section 5. Calculation of the final model accuracy is from the mean from the number of repeats.

The purpose of pruning is to minimise the risk of overfitting.The cross-validation error grows the tree to an optimal level, and the objective is to pick the tree size that minimises misclassification rate. Within R package, the printcp function provides the output for the lowest level of the rel_error can be used to calculate the optimal cp, prediction error rate in training data and prediction error rate in cross-validation (Therneau et al.; 2015).

The predicted error rate for the re-substitution cp is 48.4%, and the predicted cross-validated error rate of 56.0%. This measure is a more unbiased indicator of predictive accuracy. The tree has a misclassification rate of 56.2% in cross-validation and a predicted accuracy of 43.8%. In Figure 4 are the cp plots; the first plot the parameter value cp = 0.0001, the second plot calculated cp = 0.0012.

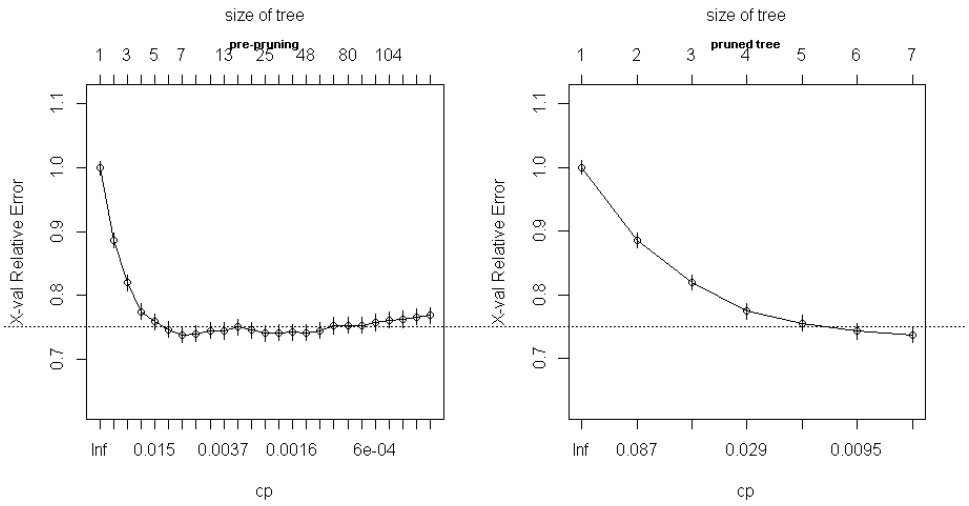

Figure 4: Complexity parameter plots of the rel_error of the pre-prune tree and pruned tree

3.7.2 Confusion Matrix and Accuracy

A confusion Matrix is used to evaluate the performance of the classifier for accuracy of classification and provide evaluation information (Sokolova and Lapalme; 2009). Accuracy is the percentage of the correct classifications with respect to the all samples. It does not say anything about the performances for negative and positive classes. Precision measures how many of the positively classified samples were really positive. There are a number of other measurements; specificity, a statistical measure of how well a binary classication test correctly identify the negative cases; true positive rate is recall the fraction of instances of a class that were correctly predicted also called sensitivity, hit rate, and recall; false positive rate is also called false alarm rate; precision is the portion of correct predictions

for a certain class or the positive predictive value; F-measure, is a combination of precision and recall. The confusion matrix can be used to compute precision and recall of a class. In a multiclass confusion tables precision and recall are calculated for all the classes, and then an overall class average is taken as a single measurement (Sokolova and Lapalme; 2009).

3.7.3 Kappa Statistic

The Kappa statistic is the measure of agreement between the predictions and the actual labels. Another interpretation is a comparison of the overall accuracy of the expected random chance accuracy. A larger Kappa metric indicates better reliability, with 0 to 0.20 as slight and bands 0.81 to 1 almost perfect (Landis and Koch; 1977). The Kappa statistic varies from 0 to 1, where 0 is the agreement equivalent to chance and 1 is perfect agreement. Mandrekar (2011) refers to the use of Kappa in clinical studies for assessing variables of interest associations and discusses the importance of evaluation and other measures in disease diagnostics. Disease prevalence can impact the results of matrices such as sensitivity, specificity, positive and negative predictive. Mandrekar (2010) reports that there is a dependency of Kappa on the prevalence of the response (chronic disease) that is the prior probability before testing.

3.7.4 Interpretability and Gini Index

Interpretability is a metric to evaluate the rule quality of a decision tree; number of rules, number of variables used in rule and the number of levels. Decision trees have two spit criteria: Gini Index(Gini Split) and Information Gain. The split method in CART algorithm is a Gini split and the split method in ID3 and C4.5 Information Gain(Entropy) which is suitable for smaller partitions. The Gini index is used to split the largest category into a separate group and each split maximizes the decrease in impurity. The response variable has 16 classes (categories), with a binary split on the variables (values; Yes and No) this is a large number of possible splits. The misclassification rate is used to measure. The decision tree comprises of; the root node and leaves (terminal or decision nodes).

3.8 Dimension reduction suitability test

3.8.1 Kaiser-Meyer-Oklin Test

An analysis test preceded the dimensional reduction stage, Kaiser-Meyer-Oklin Test (KMO) Measure of Sampling Adequacy (MSA). KMO is a measure relevant for analysing the quality of Factor Analysis. The KMO test was applied to test the variables. Beavers et al. (2013) reports that this test as useful in manual factor analysis, to assess which variables to drop from the model because they are too multicollinear. The result produced from KMO is a statistic mean measurement for each variable (Beavers et al.; 2013). The results provide an overall MSA = 0.5, hence using this test to inspect the variables did not provide values lower than 0.5. This test on inspection does not aid in the reduction of variables. The test dataset shows the same result for MSA = 0.5 with no improvement in the results.

4 Multiple Correspondence Analysis

MCA is applied using an R studio package FactoMineR by Husson, Josse, Lê and Mazet (2017). The purpose of applying MCA is to decide which variables to drop and which are the four chronic diseases that the response variable will represent. The Master list of variables are presented in Table 5. The MCA results assist in the decision process and produce information on the number of categories and the contribution details used to reduce the categories from 128 to 101.

4.1 Dimension eigenvalues and percent variance

MCA is performed on 45 variables which have a set of 128 dimensions. The results for the top five dimensions are presented in Table 7. Each eigenvalue has a variance and is calculated as a percent of the total inertia. The main purpose of the inertia is to indicate the number of axes to analysis further. The variance of the dimensions is reviewed for the viability of an cumulative variance of 98% percent. The number of dimensions that accommodate 98% cumulative variance is 101, a similar result is obtained in test dataset. The Dimensions 1, 2, 3, 4 and 5 explains a cumulative total of 17.3% of inertia in the training dataset and first five dimensions in testing dataset have an cumulative total of the 17.7% of the total inertia (Kassambara and Mundt; 2016),(Bendixen; 1996). The variables can be reduced to 101 keeping 98% viability threshold.

Table 6: First five dimensions results, cumulative variance train 17.39% and test 17.68%

	Training			Test		
Dimensions	Eigenvalue	% of variance	Cumulative % of variance	Eigenvalue	% of variance	Cumulative % of variance
dim 1	0.168	6.141	6.1410	0.168	6.213	6.213
dim 2	0.088	3.233	9.3740	0.091	3.358	9.571
dim 3	0.080	2.937	12.311	0.084	3.101	12.672
dim 4	0.075	2.754	15.065	0.069	2.563	15.235
dim 5	0.064	2.328	17.393	0.066	2.441	17.676

4.2 Coordinates, cos2 and contribution of variables

The results of MCA present coordinates, cos2 and contribution of the variable categories. The intent is to keep as many of the significant correlation coefficients as possible. The larger the percent of the variable category the more it contributes. The dimension percentage of variance and the contribution of each variable category are used to calculate the total contribution for each variable. The calculated results are sorted for importance and retention, to find the most significantly associated variables with a given principal component. The lower contribution variables are discarded (Equation 1). The squared cosine (cos2) is the quality of representation of the variable categories, and not all the points are displayed equally on the two dimensions. The cos2 measures the degree of association between categories and a particular axis.

$$TotalContribution = (Dim1\% \times Cat.Contrib1) + (Dim2\% \times Cat.Contrib2)... \quad (1)$$

The results from coordinates, cos2 and contribution of variables are visualised using scatter plots and the dimension percentages of variance presented on a scree plot (further details configuration manual). The FactoInvestigate package in R provides an automatic description of factorial analysis (Thuleau and Husson; 2017). The results provide insight into clustering of variables and the visualisation and output indicates that there are three

Table 7: Sample of variable with the calculation of the total contribution percent by category

Dimension % variance		6.141%	3.233%	2.937%	2.754%	2.328%	17.3%	
Variable.Category		Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim 1-5	
B_ALQ101.drink.12.minperyr_N	contribution	0.006	4.172	0.019	0.0260	0.010	0.783	
B_ALQ101.drink.12.minperyr_Y	contribution	0.003	2.323	0.011	0.0150	0.010	0.436	
B_ALQ101.drink.12.minperyr								1.218

clusters Figure 5. Cluster 1 has high frequency for factors questions that are 'N' and low frequency for factors that are 'Y' answers, and Cluster 2 and 3 have high and low frequency questions that have 'Y' and 'N' answers. FactoInvestigate results reports that in the analysis of a two-dimensional correspondence map that 28 axes contain important information, the relative inertia which a proportion of the total inertia of the components . This analysis provides additional information on the key dimensions and their exploratory power. The final Master dataset comprises of 28 variables which includes the class variable, 28 is a good number of variables to build the classification model.

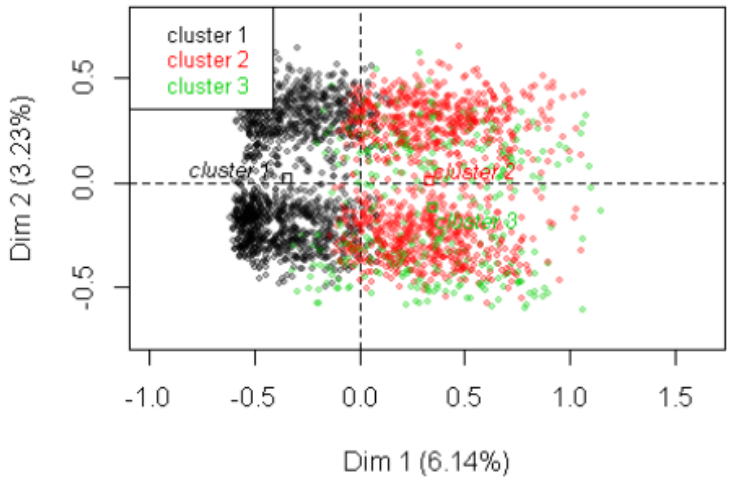

Figure 5: Visualisation plot of the three clusters observed from the output of R Studio, FactoInvestigate package.

5 Evaluation and Results

The application of the classification algorithm is by decision tree, implemented using the Rpart while Random Forest is used as a comparison for accuracy in R Studio.

5.1 Comparison results training, test and new dataset

The CART output gives results and an unweighted Kappa statistic is calculated. The overall accuracy rate is computed along with a 95 percent confidence interval and the proportion of correct predictions using test set 42.4%. The Kappa value in training is 0.334, this is between 0.21-0.40 this result is describes as 'is fair' by Landis and Koch (1977). The results for the Training and Test data (NHANES 2013–2014) and New data (NHANES 2009–2010) are presented in Table 8.

Table 8: CART results statistics

	Training	Test	New data[3]
Accuracy	0.4330	0.4244	0.4840
95% CI	(0.414, 0.4522)	(0.3945, 0.4547)	(0.4684, 0.4995)
No Information Rate	0.4114	0.2592	0.2701
P-Value ACC <NIR	0.0129	2.2e-16	2.2e-16
Kappa	0.3343	0.3125	0.3803

[3]New data is from NHANES Cycle 2009-2010.

5.2 Comparison to Random Forest

Random Forest is the average of many trees grown and produces a slightly better accuracy compared to the CART single tree (Rpart). The results from the CART model produces an overall accuracy result of 42.9% for train and test, and 48.4% new data. The overall accuracy for RF is 43.9% a slightly better result.For building the model CART is selected due to its simple and transparent decision tree.

Table 9: Random Forest results

Random forest	Training	Test	Overall Mean
Number of Trees	500	500	
Variables tried at each split	5	5	
Accuracy	43.57%	44.23%	43.89%
OOB estimate of error rate	56.43%	55.77%	56.98%

5.2.1 Decision Tree visualisation

The resultant model separated the classes into eight groups and not all of the 27 variables appear in the final mode. The variables of importance is the list in CART are the same variables as the decision tree Figure 6. The root (M_BPQ040) represents the attribute that plays a central role in classification and the leaf represents the predicted class. The numbers at the bottom of the terminal branches indicate the probability in each data subset. Figure 6 is based on the decision tree from the Rpart model.

5.3 Discussion

In this research, cross-validation is used to produce the results. The cross-validation relies on some independently selected subset of data (test data). The training model is fine-tuned using the k-fold method (k=10). The aim is to have a small tree, a tree with least cross-validated error to avoid any over-fitting of the data. The accuracy between the cross-validation method and pre-processing is comparable at 43.2% to the pre-processing substitution cp. The accuracy of RF is just slightly better than Rpart model. The comparisons results indicate that overfitting is not an issue and there is some reliability in the accuracy reported. There are eight rules at deployment stage of the application - this small given the size of the data and the numbers of classes.

The CART solution offers a slightly lower accuracy compared to RF. However, accuracy is not the only measurement for evaluating. The statistic results CART are in Table 8. The decision tree offers a set of rules in the format if-else and indicates that the sample belongs to a certain class. RF is a black box method and gives insight into variable important. Transparency and traceability are important in the creation of a final set of rules for the application the CART model is the preferred choice.

Statistical results from the CART model includes the matrices by class: specificity, sensitivity and prevalence. Six of sixteen potential classifiers have values for all three

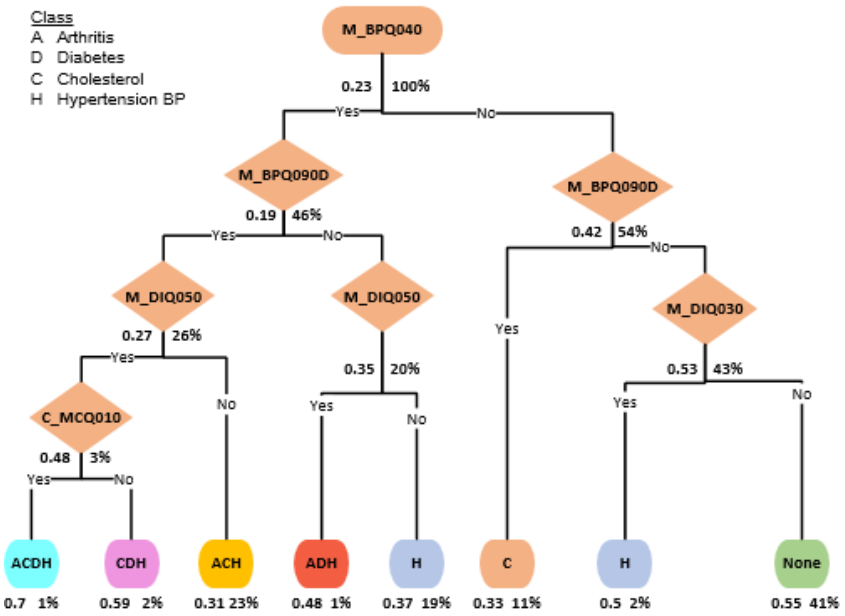

Figure 6: Decision tree with probability of the class

matrices are; ACH, ADH, AH, CDH, H and None. Classes None and ADH have the two highest sensitivity, with true positive results of None (54.9%) and ADH (48.4%). All classes have results for specificity (true negative) ranging from 91.1% to 99.4%. There is a likely-hood of both the probability of type I error (false positive) and type II error (false negative) among the classes. The two classes None (77.1%) and ADH (73.5%) have the highest accuracy percentages. The Kappa statistic is used to measure agreement between the outcomes. The Kappa ranges from 0.313 to 0.380, all between 0.21-0.40 this result 'is fair'. CART rules are easy to implement compared to RF.

6 Implementation

Described in this section is the implementation applied for the data mining process and the Machine learning algorithms.

6.1 Applications Environment

The applications and environments for this research are reported in the configuration manual. The main technologies are Microsoft SQL Server Management Studio 2014, R Studio, Microsoft Access and SPSS. Within R there are many packages available for the creation of decision trees, and rpart from the Caret package was chosen to create a decision tree. Alternative options are available using R packages for decision trees such as Random Forest and rtree (Breiman and Cutler; 2016), (Therneau et al.; 2015)), (Kuhn et al.; 2016),(Liaw and Wiener; 2002), (Thuleau and Husson; 2017), (Kassambara and Mundt; 2016).

6.2 Building an application

Data mining techniques are used to extract insight from large amounts of data.The techniques include dimensional reduction and prediction classification. The objective is to find the presence or absence of a characteristic - in this scenario, a chronic disease. Decision tree algorithms are the chosen predictive classification models for this research. The decision tree algorithm approach is to build the tree and apply the model. The response variable has multiple categories, so a standard classification, where the response variable only has two categories, is not sufficient.

The design of the build is; train phase, test and validate and an application build phase. The approach in this research is to reduce the number of variables and build a decision tree using the variables that contribute the best predictive power. The extracted decision tree rules are re-encoded into an SQL stored procedure syntax in a SQL database.

6.3 Deployment

The purpose of the deployment phase is to verify the parameters and apply the model. An objective of the research is to build a classifier application with a user interface, named as MCDSAT. This application is in the format of a questionnaire and develops in Microsoft Access. The purpose is to test, simulate end users query with the 'unseen' dataset. To record results for the classifier, which comprises of the four chronic diseases in this study arthritis (A), diabetics (D), cholesterol (C) and hypertension (H), diseases or for a classifier of none. The 'unseen' dataset query is sourced from the NHANES Questionnaires cycle 2009-2010. The final stage of implementation builds the discovered decision tree rules into an application, in this research a SQL database and a user interface MS Access. The final stage of to test the working prototype, use individual queries and record results from a working prototype.

7 Conclusion and Future Work

This paper presents the research undertaken to develop an MCDSAT including the approach of dimensional reduction of the categorical variable set of survey questions and the application of classification algorithms for predicting the class.

The objective of the research is to predict the classification of a person at risk of a single or multiple chronic diseases using questionnaire data and evaluate the performance. The objective attained, and the results in Section 5. The two main challenges are the reduction of the data and the choice of classification. A strength of the research is that it does achieve the deliverable - the development of a simple application tool and the evaluation of the model. The model could also apply to other business domains where survey questionnaires are a tool to collect information, such as marketing, banking or customer support. The CART method gives a set of rules that are interpretable as opposed to the slightly more accurate RF. There is apparent value in developing a high accuracy tool which can screen or diagnosis of a chronic disease. A limitation of the results is that they do not achieve an accuracy level that is enough for a chronic disease diagnosis or progress pre-screening.

The recommendations for future work include the study of other components from the NHANES dataset such as dietary and physical exercise. There is the potential to progress the model process by further reduction of the model variables and compare the

results for accuracy. Also, the interrogate of repeating the classifier models for a single chronic disease and compare results to decide if there is an improvement. In this research, MCA applied for data reduction method on the health questionnaires (NHANES). The MCA method can provide insight into associations, clustering in sights and variety of visualisations of results (Thuleau and Husson; 2017), (Kassambara and Mundt; 2016).

Acknowledgements

My sincere gratitude to my supervisor Barry Haycock for his continuous support, motivation, enthusiasm and immense knowledge during this M.Sc research project.

References

Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., Bahadorian, B. and Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease, *Computer methods and programs in biomedicine* **111**(1): 52–61.

American Council on Exercise (2017). Risk Assessment Calculator.
URL: *https://www.acefitness.org/education-and-resources/lifestyle/tools-calculators/risk-assessment-heart-attack*

American Diabetes Association (2017). Type 2 Diabetes Risk Test.
URL: *http://www.diabetes.org/are-you-at-risk/diabetes-risk-test/*

American Diabetes Association and others (2004). Screening for type 2 diabetes, *Diabetes care* **27**(suppl 1): S11–S14.

Azevedo, A. I. R. L. and Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview, *IADS-DM* pp. 182–185.

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J. and Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research, *Practical assessment, research & evaluation* **18**.

Bendixen, M. (1996). A practical guide to the use of correspondence analysis in marketing research, *Marketing Research On-Line* **1**: 16–36.

Breiman, L. and Cutler, A. (2016). Breiman and Cutlers Random Forests for Classification and Regression.
URL: *https://www.stat.berkeley.edu/ breiman/RandomForests/*

Breiman, L., Friedman, J. and Olshen, R. (1985). Classification and regression trees, *Wadsworth International Group* **8**: 452–456.

Centers for Disease Control and Prevention (2017). NHANES 2013-2014 Survey Questionnaires.
URL: *http://bit.ly/2nhfMeE*

Chaurasia, V. and Pal, S. (2013). Data mining approach to detect heart diseases, *International Journal of Advanced Computer Science and Information Technology (IJACSIT)* **2**(4): 56–66.

Conway, J. M. and Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research, *Organizational research methods* **6**(2): 147–168.

Costa, P. S., Santos, N. C., Cunha, P., Cotter, J. and Sousa, N. (2013). The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing, *Journal of aging research* **2013**: 302163.
URL: *http://europepmc.org/articles/PMC3810057*

Forouzanfar, M., Afshin, A., Alexander, L.T.and Anderson, H., Bhutta, Z., Biryukov, S., Brauer, M., Burnett, R., Cercy, K., Charlson, F.J. and Cohen, A. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease study 2015, *The Lancet* **388**(10053): 1659–1724.

Gellish, R. L., Goslin, B. R., Olson, R. E., McDonald, A., Russi, G. D. and Moudgil, V. K. (2007). Longitudinal modeling of the relationship between age and maximal heart rate, *Medicine and science in sports and exercise* **39**(5): 822–829.

Harvard T.H. Chan School of Public Health (2017). Healthy Heart Score.
URL: *https://healthyheartscore.sph.harvard.edu/*

Heikes, K. E., Eddy, D. M., Arondekar, B. and Schlessinger, L. (2008). Diabetes risk calculator, *Diabetes care* **31**(5): 1040–1045.

Husson, F., Josse, J., Lê, S. and Mazet, J. (2017). Package FactoMineR.
URL: *http://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf*

Husson, F., Lê, S. and Pagès, J. (2017). *Exploratory multivariate analysis by example using R*, CRC press.

Josse, J. and Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations, *Computational Statistics & Data Analysis* **56**(6): 1869–1879.

Kassambara, A. and Mundt, F. (2016). Package factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y. and Candan, C. (2016). *caret: Classification and Regression Training*. R package version 6.0-68.
URL: *http://CRAN.R-project.org/package=caret*

Kumari, M., Vohra, R. and Arora, A. (2014). Prediction of Diabetes Using Bayesian Network, *International Journal of Computer Science and Information Technologies (IJCSIT)* **5**(4): 5174–5178.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data, *biometrics* pp. 159–174.

Lê, S., Josse, J., Husson, F. et al. (2008). FactoMineR: an R package for multivariate analysis, *Journal of statistical software* **25**(1): 1–18.

Li, T., Zhang, C. and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics* **20**(15): 2429–2437.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest, *2* **3**: 18–22.
URL: *URL http://CRAN.R-project.org/doc/Rnews/*

Liu, S., Dissanayake, S., Patel, S., Dang, X., Mlsna, T., Chen, Y. and Wilkins, D. (2014). Learning accurate and interpretable models based on regularized random forests regression, *BMC systems biology* **8**(S3): S5.

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

Mandrekar, J. N. (2010). Simple statistical measures for diagnostic accuracy assessment, *Journal of Thoracic Oncology* **5**(6): 763 – 764.
URL: *http://www.sciencedirect.com/science/article/pii/S1556086415305013*

Mandrekar, J. N. (2011). Measures of interrater agreement, *Journal of Thoracic Oncology* **6**(1): 6–7.

National Heart Foundation of Australia (2017). Target Heart Rate Calculator.
URL: *http://www.heartonline.org.au/resources/calculators/target-heart-rate-calculator*

National Heart, Lung, and Blood Institute and others (2013). Assessing Cardiovascular Risk: Systematic Evidence Review From the Risk Assessment Work Group, *Bethesda, MD: National Institutes of Health* .

Polat, K. and Güneş, S. (2007). A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and airs, *Computer methods and programs in biomedicine* **88**(2): 164–174.

Sathyadevi, G. (2011). Application of CART algorithm in hepatitis disease diagnosis, *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on*, IEEE, pp. 1283–1287.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks, *Information Processing & Management* **45**(4): 427–437.

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunananthan, S., Bandeen-Roche, K., Béland, F. and Bergman, H. (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables, *Journal of clinical epidemiology* **63**(6): 638–646.

Therneau, T., Atkinson, B. and Ripley, B. (2015). *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-10.
URL: *http://CRAN.R-project.org/package=rpart*

Thuleau, S. and Husson, F. (2017). Factoinvestigate: Automatic description of factorial analysis. R package version 1.1.
URL: *https://CRAN.R-project.org/package=FactoInvestigate*

Tuszynski, J. (2014). *caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.* R package version 1.17.1.
URL: *https://CRAN.R-project.org/package=caTools*

UCLA:Statistical Consulting Group (2011). R Library Contrast Coding Systems for categorical variables.
URL: *https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/*

World Health Organization (2015). Media centre Noncommuncable Diseases Fact Sheet.
URL: *http://www.who.int/mediacentre/factsheets/fs355/en/*

World Health Organization and others (1986). The Ottawa charter for health promotion: first international conference on health promotion, Ottawa, 21 November 1986, *Geneva:WHO* .