Dublin City University

Use of Machine Learning Technology in the Diagnosis of Alzheimer's Disease

Author:
Noel O'Kelly

Supervisors:
Prof. Alan F. Smeaton and Dr. Kate Irving

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in the

School of Computing

September 2016

Declaration of Authorship

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Master of Science is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____

(Candidate) ID: 52178340

Date: _____

Acknowledgements

I would firstly like to take this opportunity to thank my supervisors, Prof. Alan Smeaton and Dr. Kate Irving, for their guidance and expertise in development of this work, Prof. Smeaton for his expertise in machine learning and data analytics, and Dr. Irving for her support and advice on the subject of Alzheimer's disease. I would also like to thank my colleagues Jim O'Donoghue and Micheal Scriney for their assistance with technical issues and advice. Special thanks are due to Owen Corrigan at the Insight Centre for Data Analytics in DCU for his help setting up and assistance with the Machine Learning algorithms and to my good friend Colette Boland for her positive encouragement and proofreading. Finally I would like to thank my daughter and son, Francisca and Brian O'Kelly, for sharing their experiences of study at this level, and in Francisca's case, for proofreading the thesis.

The work reported in this thesis was partially supported by the Elevator project and by Science Foundation Ireland under grant number SFI/12/RC/2289 (Insight Centre).

Special thanks is due to the National Alzheimer's Coordinating Center (NACC) at the University of Washington in the USA, who provided the dataset used in our work and which was invaluable to us.

Contents

List of Figures

List of Tables

Abstract

Use of Machine Learning Technology in the Diagnosis of Alzheimer's Disease

by Noel O'Kelly

Alzheimer's disease (AD) is thought to be the most common cause of dementia and it is estimated that only 1-in-4 people with Alzheimer's are correctly diagnosed in a timely fashion. While no definitive cure is available, when the impairment is still mild the symptoms can be managed and treatment is most effective when it is started before significant downstream damage occurs, i.e., at the stage of mild cognitive impairment (MCI) or even earlier. AD is clinically diagnosed by physical and neurological examination, and through neuropsychological and cognitive tests. There is a need to develop better diagnostic tools, which is what this thesis addresses.

Dublin City University School of Nursing and Human Sciences runs a memory clinic, Memory Works where subjects concerned about possible dementia come to seek clarity. Data collected at interview is recorded and one aim of the work in this thesis is to explore the use of machine learning techniques to generate a classifier that can assist in screening new individuals for different stages of AD. However, initial analysis of the features stored in the Memory Works database indicated that there is an insufficient number of instances available (about 120 at this time) to train a machine learning model to accurately predict AD stage on new test cases.

The National Azheimers Cordinating Center (NACC) in the U.S collects data from National Institute for Aging (NIA)-funded Alzheimer's Disease Centers (ADCs) and maintains a large database of standardized clinical and neuropathological research data from these ADCs. NACC data are freely available to researchers and we have been given access to 105,000 records from the NACC. We propose to use this dataset to test the hypothesis that a machine learning classifier can be generated to predict the dementia status for new, previously unseen subjects. We will also, by experiment, establish both the minimum number of instances required and the most important features from assessment interviews, to use for this prediction.

Chapter 1

Alzheimer's Disease: An Overview

1.1 Introduction

Alzheimer's Disease (AD) is a degenerative brain disease that effects humans and is thought to be the most common cause of dementia, though dementia can also be caused by other diseases and conditions. It is characterized by a decline in memory, ability to formulate and use language, problem-solving and other cognitive skills and these characteristics affect a person's ability to perform everyday activities. This decline in human abilities occurs because nerve cells (neurons) in the parts of the brain involved in cognitive function have been damaged and no longer function normally. In Alzheimer's disease, neuronal damage eventually affects parts of the brain that enable a person to carry out basic bodily functions such as walking and swallowing. Alzheimer's Disease is a terminal disease with no disease-modifying treatment available as yet. Dementia is an umbrella term which is used to describe a set of symptoms, and there are many different types of dementia including Alzheimer's Disease, vascular dementia, dementia with Lewy bodies, and others, but dementia of the Alzheimer's type (AD) is by far the most common cause of dementia, and this is the type of dementia this thesis is concerned with.

According to the Alzheimer's Association's 2015 report [2], an estimated 5.3 million US citizens have Alzheimer's disease. While 5.1 million of these are aged more than 65 years,

approximately 200,000 are aged less than 65 years and have what is called younger onset of Alzheimer's Disease. By the middle of this century, the number of people living with AD in the United States is projected to grow by nearly 10 million, fuelled in large part by the aging baby boom generation. Today, someone in the USA develops AD every 67 seconds. By 2050, one new case of AD is expected to develop every 33 seconds, resulting in nearly 1 million new cases per year, and the estimated prevalence is expected to range from 11 million to 16 million. In 2013, official death certificates in the United States recorded 84,767 deaths from AD, making AD the sixth leading cause of death in the United States and the fifth leading cause of death in Americans aged 65 years or greater. Between 2000 and 2013, deaths resulting from heart disease, stroke and prostate cancer decreased by 14%, 23% and 11%, respectively, whereas deaths from AD increased by 71%.

The figures mentioned above are for the United States alone. Worldwide, nearly 44 million people have Alzheimer's or a related form of dementia.
Predictions from Ireland show a similar growth pattern. The Irish National dementia Strategy, Published by the Department of Health in December 2014, contained estimates for the incidence of AD for the years 2011 - 2046 in the Republic of Ireland. The estimates are that the number of sufferers in total for all age groups will increase from 47, 829 in 2011 to a total (all age groups) of 152,157 in 2046. In percentage terms, this is greater than the predicted growth in numbers for the US.

Only 1-in-4 people with Alzheimer's disease have been diagnosed, according to Alzheimer's Disease International [2]. Alzheimer's and dementia is most common in Western Europe (North America is close behind) and while no definitive cure for Alzheimer's Disease is available, suffering can be lessened by compassionate skilled post-diagnosis support.

Alzheimer's Disease is a serious personal, medical and social problem. Recent research indicates early and accurate diagnosis as the key to effectively coping with it. However, even in the later stages of the disease, diagnosis is inaccurate 50% of the time according to Boise *et al.* [3]. Even when the disease is diagnosed correctly, monitoring the progression of the disease over time is costly. Treatment is thought to be most effective when it is started before significant downstream damage occurs, i.e. before clinical diagnosis of Alzheimer's Disease, at the earlier stage of mild cognitive impairment (MCI) or even earlier.

It is widely accepted that an early detection of dementia can lead to a more effective intervention and the limiting of morbidity (Petersen *et al.* [4]). Furthermore, Petersen *et al.* [5] conclude from their work that people who meet the criteria for MCI can be differentiated from healthy control subjects and those with very mild Alzheimer's Disease. This group of subjects appear to constitute a clinical entity that can be characterized for treatment interventions.

To date, the diagnosis of most forms of mental disorder has been based on clinical observation. Specifically these include the identification of symptoms that tend to cluster together, the timing of the symptoms' appearance, and their tendency to resolve, recur or become chronic. There is currently no cure for Alzheimer's Disease and we lack any form of reliable and effective early diagnostic tools. Boise *et al.* [3] confirm earlier studies regarding low rates of clinical assessment and diagnosis and postulate a possible explanation for this in the subtlety of dementia symptoms combined with the constraints physicians face in their clinical practice. Alzheimer's Disease is clinically diagnosed by performing physical and neurological examinations, and checking other signs of intellectual impairment through standard neuropsychological and cognitive tests. The general approach is based around *diagnosis by elimination*, i.e. ruling everything else out until Alzheimer's Disease remains the last option.

In addition to the above clinical measures, according to Dubois *et al.* [6] the guidelines for the diagnosis of Alzheimer's Disease emphasise the role that can be played by using various biomarkers. These include measures from magnetic resonance imaging (MRI), positron emission tomography (PET), cerebrospinal fluid (CSF) protein profiles as well as analysis of genetic risk profiles though these are expensive and difficult to scale to large numbers of assessments. Clearly, there is a need to develop better diagnostic tools for Alzheimer's Disease diagnosis, possibly using data mining and data analysis techniques, which is what we explore in this thesis. If new drugs or prevention strategies were developed and proven to be effective, then an early diagnosis might enable intervention at an earlier stage which would be of proven benefit, yet we are still at a time when clinical diagnosis is carried out using only the signs and symptoms of the disease, and this is challenging.

There is no single test that can show whether a person has or does not have Alzheimer's Disease. While physicians can almost always determine if a person has dementia, it

may be difficult to determine the exact cause. Diagnosing Alzheimer's requires careful medical evaluation, including:

- A thorough assessment of medical and family history;
- Input from a family member or persons close to the individual about changes in their cognitive skills or behaviour;
- Mental status testing;
- A physical and neurological examination;
- Tests (such as blood tests and/or brain imaging) to rule out other causes of dementia-like symptoms such as a tumour that could explain the individual's symptoms.

It is by aggregating the outputs from the above set of assessments that a physician can make a diagnosis of Alzheimer's Disease and this requires skill, expertise, and experience., none of which can be easily replicated.

1.2 The Elevator Project and Memory Works

The Elevator Project is a programme developed by Dublin City University in collaboration with the Health Services Executive (HSE) and supported by Atlantic Philanthropies. It is an education and empowerment programme to help individuals, communities and health systems engage appropriately with people with dementia. Following a needs analysis of the dementia-related education that is required across the community, carried out in 2014 and reported in [7], the Elevator work programme has been developed into a multi-faceted approach to education, and consists of the following elements:

- Dementia champions training for dementia care;
- Training in everyday ethical decisions for family and health professionals;
- Mechanisms to raise dementia awareness;
- Dedicated dementia training for GPs, etc.;

- Training in dementia awareness;
- Training in psychosocial skills;
- Memory assessment and the development of online tools for health and social care professionals.

One part of the of delivery of the Elevator project is to improve assessment and integration of such assessment into everyday practice, which are now described below.

The School of Nursing and Human Sciences at Dublin City University runs the "Memory Works" which has been in operation for several years. Memory Works is a screening clinic aimed at identifying people with a pathological reason for their memory problems and those who do not. Its aim is to fill a gap in the existing health service for people who feel that their memory is a problem. The service, which is available to anyone over 40, years of age and works on a self-referral or GP-referral basis.

Poor memory can be the result of many things, including lack of exercise, mental stimulation, emotional worries, a stressful lifestyle, problems at home or at work, etc. In short, what the clinic aims to do is to help alleviate clients' concerns about their poor memory by helping them to find out the underlying cause of their problem.

This project supports the second work stream of this study in that it attempts to take some of the learning from the Memory Works clinic and converts it into new knowledge, tools and educational materials for supporting memory assessment.

1.3 The Elevator Project Dataset

One feature of the Elevator project is the creation and management of records relating to visits by people to the Memory Works memory clinic. An online records-management system has been created where clinicians sign into the system as a member of a clinic and can then add, update and remove patient records from the System. The data is entered through a web application which consists of a number of forms or screens, each relating to a particular group of information e.g. patient history, patient lifestyle, etc. The data is subsequently stored in a database in the cloud, using the Google App Engine Cloud SQL relational database.

With such a resource available, this thesis sets out to explore ways in which the data collected in such a system might be used to help future patients, perhaps by using machine learning techniques to learn patterns or build models which can be used to assist with the screening of future patients visiting the clinic, for assessment for different stages of Alzheimer's Disease. However, in order to do this there is a need to establish the number of instances we need in order to be statistically reliable with such automated or semi-automated, or computer-assisted screening, as well as determining which are the most important features from a patient visit which are present in the instances in order to achieve a good degree of accuracy for such screening assistance. We shall return to this point later.

Figure 1.1 shows a graphical outline of the database schema used in the relational database underpinning the online system. There are multiple patients, and each patient might have multiple forms, each from relating to a data grouping from an input screen.

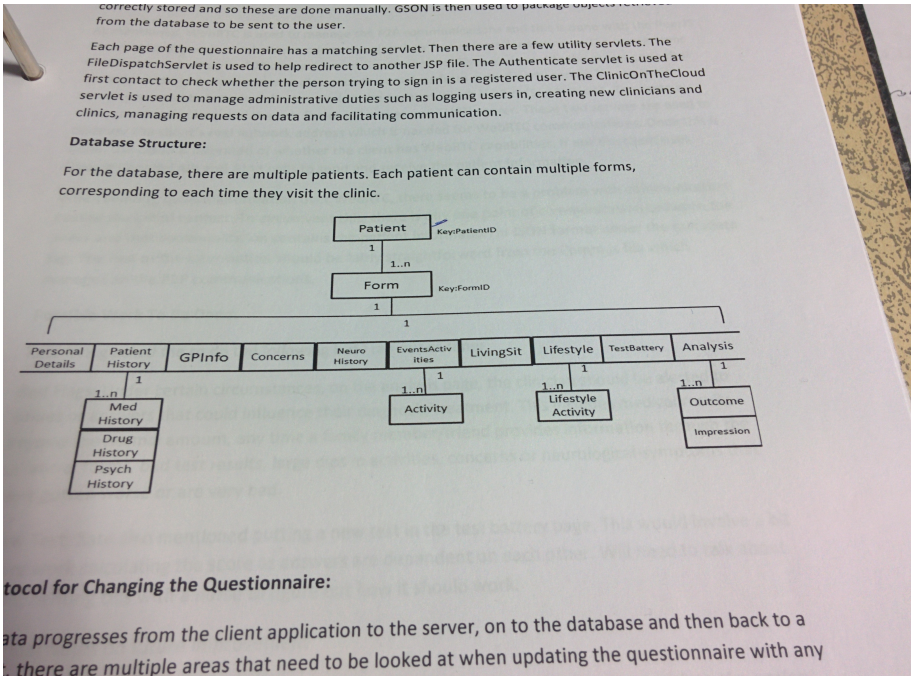

FIGURE 1.1: Database schema for Elevator data

Each of the forms in the Elevator project dataset correspond to each time a patient visits the clinic. The forms collect information on a variety of areas of a patient, including:

- Personal details

- Patient's medical history
- GPInf - information on the referring GP, if there is one
- Patient concerns
- Neuro history
- Patient activities
- Living situation
- Lifestyle
- Neuro-psychological test battery (e.g. MMSE)
- Analysis (clinical diagnosis)

Before going further into the idea of using machine learning or data mining or other forms of data processing to assist with the screening process, we will briefly summarise some of the related work where such data processing techniques have been used in diagnosis or screening of Alzheimer's Disease.

1.4 The Nature of Alzheimer's Disease Diagnosis

Data mining and big data analytics can be used to provide medical informatics researchers and practitioners with systematic technical solutions for the analysis of large volumes of medical data. Such analysis can lead to the construction of predictive models, including addressing the goals of diagnosing, treating, helping, and healing patients with mental health disorders such as Alzheimer's Disease. Diagnostic criteria for mental health diseases are frequently based on clinical and psychometric assessment and it is the data from these assessments that forms the basis for the data analytics techniques that are explored in this thesis.

A 2012 literature review by Yoo *et al.* [8] established that data mining techniques are being successfully employed in the healthcare field for a multitude of purposes, including prediction of healthcare costs, disease diagnosis/prognosis and the discovery of hidden biomedical and healthcare patterns from related databases. That paper describes how

data mining technologies have been used and reports that classification is the core data mining method used in bioinformatics and in biomedicine. It concludes that data mining has been successfully and widely used in these fields and describes some of the problems that hamper the clinical use of data mining by health professionals.

The National Institute on Aging and the Alzheimer's Association charged a working group with the task of revising the 1984 criteria for diagnosing Alzheimer's Disease dementia. Revised criteria and guidelines for diagnosing Alzheimer's Disease were proposed and published in 2011 [9].The most significant aspect of the new criteria was the first introduction of the presence or absence of biomarkers into the core diagnostic framework. The workgroup sought to ensure that the revised criteria would be flexible enough to be used by both general healthcare providers without access to neuropsychological testing, advanced imaging, or cerebrospinal fluid measures, techniques which we describe later. They also wished to include that it could be used by specialized investigators involved in research or in clinical trial studies who would have these tools available. They presented criteria both for all-cause dementia and for Alzheimer's Disease dementia. The working group retained the general framework of probable Alzheimer's Disease dementia from the 1984 criteria. On the basis of the subsequent 27 years of experience, they made several changes in the clinical criteria for the diagnosis. They also retained the term "possible Alzheimer's Diseas dementia", but redefined it in a manner more focused than before. Biomarker evidence was also integrated into the diagnostic formulations for probable and possible Alzheimer's Disease dementia for use in research settings.

While the core clinical criteria for Alzheimer's Disease dementia will continue to be the cornerstone of the diagnosis in clinical practice, biomarker evidence is expected to enhance the pathophysiological specificity of the diagnosis of Alzheimer's Disease dementia. Much work lies ahead for validating the biomarker diagnosis of Alzheimer's Disease dementia and they recommend that Alzheimer's be considered a slowly progressive brain disease that begins well before clinical symptoms emerge.

From a biomarker point of view, the hallmark pathologies of Alzheimer's Disease are the progressive accumulation of the protein fragment beta-amyloid (plaques) outside neurons in the brain, and the presence of twisted strands of the protein tau (tangles) inside neurons. These changes are eventually accompanied by the damage to, and death

of, neurons (Alzheimer's Association) [2]. An autopsy to detect these biomarkers is regarded as the gold standard for the detection of Alzheimer's Disease in a subject.

1.5 Summary

This thesis sets out to examine how data analytics, and in particular machine learning, can or could be used to help with assessment of people. In Chapter 2 we review the literature concerned with the extensive research already carried out over the last few years into how biomarker and imaging technology can be put to use to assist in implementing cost effective diagnostic tools for Alzheimer's Disease diagnosis. These are complex and the need for expertise makes them expensive. In Chapter 3 we then describe, at a high level, the topic of machine learning covering how it works, and the software tools that are available. As a worked example, we apply some machine learning software to some data from the Memory Works clinic introduced earlier, and from this we then present the Main Research Question and a number of research sub questions which are the basis for this thesis.

In Chapter 4 we then introduce a second dataset of patient data that we acquired for this work which we refer to as the NACC (National Alzheimer's Coordinating Center) data, from the US. We also describe the pre-processing we did on this dataset to prepare it for machine learning based processing.

Chapter 5 describes the steps we followed in carrying out an initial set of experiments on the NACC dataset, and then we present the results of our machine learning experiments on predicting outcome, and the number of viable subjects needed for reliable classification. Throughout this chapter we address most of our research sub questions, raised earlier in the thesis.

In the final Chapter, Chapter 6, we revisit our research question and the set of research sub questions to see have they been answered satisfactorily and we discuss potential future work.

Chapter 2

Literature Review: Alzheimer's Disease Diagnosis

This chapter sets out to gather and then review research into the development of the tools required for an objective diagnosis of Alzheimer's Disease. It also includes a brief description of the theory underlining Machine Learning techniques.

The National Institute on Aging and the Alzheimer's Association charged a work/group with the task of revising the 1984 criteria for diagnosing Alzheimer's Disease dementia. Revised criteria and guidelines for diagnosing Alzheimer's Disease were proposed and published in 2011 [9]. The most significant aspect of the new criteria was the first introduction of the presence or absence of biomarkers into the core diagnostic framework. The workgroup sought to ensure that the revised criteria would be flexible enough to be used by both general healthcare providers without access to neuropsychological testing, advanced imaging, or cerebrospinal fluid measures. They also wished to include that it could be used by specialized investigators involved in research or in clinical trial studies who would have these tools available. They presented criteria both for all-cause dementia and for Alzheimer's Disease dementia. The work/group retained the general framework of probable Alzheimer's Disease dementia from the 1984 criteria. On the basis of the subsequent 27 years of experience, they made several changes in the clinical criteria for the diagnosis. They also retained the term "possible Alzheimer's Disease dementia", but redefined it in a manner more focused than before. Biomarker evidence was also integrated into the diagnostic formulations for probable and possible

Alzheimer's Disease dementia for use in research settings. The core clinical criteria for Alzheimer's Disease dementia will continue to be the cornerstone of the diagnosis in clinical practice, but biomarker evidence is expected to enhance the pathophysiological specificity of the diagnosis of Alzheimer's Disease dementia. Much work lies ahead for validating the biomarker diagnosis of Alzheimer's Disease. They recommend that Alzheimer's be considered a slowly progressive brain disease that begins well before clinical symptoms emerge.

For our literature survey and review for this thesis, Web of Science (WoS) was searched for papers published in the years 2004-2015 to find recent papers concerning the use of data mining/machine learning in healthcare, specifically in the diagnosis of Alzheimer's Disease and it's precursor, Mild Cognitive Impairment (MCI). We attempt to establish what approaches researchers are using to harness this form of data analysis technology to assist in the diagnosis of Alzheimer's Disease. The search years were chosen to ascertain what was the focus of research both before and after the publication of the recommendations from the workgroup established to revise the criteria for diagnosis of Alzheimer's Disease MCKhann *et al.* [9]. Some earlier papers were included in our analysis because they were referenced in the papers found by the initial WoS search.

A search with the keywords "diagnosis, dementia and machine learning" yielded 92 papers published during that period. Scanning the papers and reviewing their abstracts revealed that most of them were concerned with the use of neuroimaging, EEG or PET scan modalities for diagnosis. There are also indications that speech analysis is popular in the research literature and is being used as a less costly tool in diagnosis. The full texts of research papers were included for deeper reading and analysis after reading their abstracts to establish their relevance.

In related work, data mining and Big Data analytics provides medical informatics researchers and practitioners with systematic technical solutions for the analysis of medical data and the construction of predictive models, including the goal of diagnosing, treating, helping, and healing patients with Mental Health disorders such as Alzheimer's Disease. Diagnostic criteria for Mental Health diseases are frequently based on clinical and psychometric assessment.

A 2012 literature review by Yoo *et al.* [8] established that data mining techniques are being successfully employed in the healthcare field for a multitude of purposes, including

prediction of healthcare costs, disease diagnosis/prognosis and the discovery of hidden biomedical and healthcare patterns from related databases. It describes how data mining technologies have been used and reports that supervised machine learning classification techniques are the core data mining method used in bioinformatics and biomedicine. It concludes that data mining has been successfully and widely used in these fields and describes some of the problems that hamper the clinical use of data mining by health professionals.

Perhaps because of the recent availability of large datasets such as that provided by the Alzheimer's Disease Neuroimaging Inititive (ADNI) [10] and the 2011 revision of the diagnostic criteria described earlier, a considerable amount of current research in the diagnosis of Alzheimer's Disease, and it's precursor, (MCI) is focused on the use of neuroimaging to detect the known biomakers associated with Alzheimer's Disease and MCI. Many research initiatives to address this problem are ongoing worldwide, frequently centered around the ADNI and it's datasets.

A review of all papers published since the inception of the initiative [11] reports that approximately 500 papers have been published as a direct result of releasing ADNI to researchers, to the end of 2013. For more up to date information see `http://www.adni-info.org`

Overall, from a high level of our literature review, it was found that the published research papers in this area tended to focus on two main areas of research: biomarkers and neuroimaging, but with an increasing interest in speech analysis, and in the subsections below we shall address each of these.

2.1 Biomarkers and Neuroimaging for Alzheimer's Disease Diagnosis

From a biomarker point of view, the hallmark pathologies of Alzheimer's Disease are the progressive accumulation of the protein fragment beta-amyloid (plaques) outside neurons in the brain, and the presence of twisted strands of the protein tau (tangles) inside neurons. These changes are eventually accompanied by the damage to, and death of, neurons. (Alzheimer's Association) [2].

A position paper by Dubois *et al.* [6] considered the strengths and the limitations of the workgroup diagnostic criteria proposals. It proposes that topographical biomarkers of the disease such as volumetric MRI and flourodeoxyglucose PET might better serve in the measurement and monitoring the course of the disease.

Traditionally, the clinical diagnosis of dementia has focused on clinical assessment, neuropsychological testing, and the exclusion of other possible causes. As we saw earlier, the National Institute of Ageing and the Alzheimer's Association have issued new diagnostic criteria for Alzheimer's Disease and MCI, and now suggest the use of two other complementary modalities, cerebrospinalfluid (CSF) biomarkers and neuroimaging [9], [12].

Magnetic resonance imaging (MRI), nuclear magnetic resonance imaging (nMRI), functional magnetic resonance imaging (fMRI) or magnetic resonance tomography (MRT) are each forms of medical imaging techniques used in radiology to investigate the anatomy and physiology of the body in both health and disease. MRI scanners use magnetic fields and radio waves to form images of the body. The technique is widely used in hospitals for medical diagnosis, staging of disease and follow-up without exposure to ionizing radiation.

Querbes *et al.* [13] noted that "brain atrophy measured by magnetic resonance structural imaging has been proposed as a surrogate marker for the early diagnosis of Alzheimer's Disease." Specifically, they suggest that the thickness of the cortex in the brain is a biomarker for the presence of, or a predictor of, Alzheimer's Disease. Figure 2.1 below is supportive of this proposal and shows time lapse brain scans with healthy brain activity shown in the red and blue areas and rapidly spreading areas of cell death (gray areas) in a subject with Alzheimer's Disease. About 5% of brain cells die each year in someone with Alzheimer's, compared to less than 1% in a senior who is aging normally.

Other biomarkers under investigation for Alzheimer's Disease diagnosis using MRI scanning include the build-up, or dissipation, of a protein called beta-amyloid in the living brain. To help with the detection of this, CSF total protein is a test used to determine the overall amount of protein in spinal fluid, also called cerebrospinal fluid (CSF).

A study by Hansson *et al.* [14] shows that CSF measures can be used to predict AD, with a sensitivity of 93% and a specificity of 83% for detection of incipient Alheimer's

FIGURE 2.1: Damage to the brain of a subject with Alzheimer's Disease

Disease in patients with MCI. However, the study rightfully did mention some of the difficulties with CSF measurement including site-to-site variation in assay results and no clear agreement in cut-off values. Also the process of taking a sample of CSF is invasive, requiring a lumbar puncture (also known as a spinal tap) which is both dangerous as it can lead to infection, as well as being uncomfortable and stressful for the patient.

Adaszewski *et al.* report in [15] that MRI is well established as a non-invasive technique for detecting biomarkers for Alzheimer's Disease. The techniques they describe demonstrate the potential for reliable early diagnosis. They report that using both Structural Magnetic Resonance Imaging (sMRI) and machine learning methods, there is evidence of sufficient accuracy in discriminating between Alzheimer's disease patients from not only healthy controls but also from other common types of dementia. In the case of early Alzheimer's Disease detection, a form of machine learning using a Support Vector Machine (SVM) classifier technique, (which we address later in the thesis), convincingly demonstrates the potential for reliable early diagnosis.

In more recent developments, Casanova *et el.* [16] introduced new metrics for assessing Alzheimer's Disease risk based on Structural Magnetic Resonance Imaging (sMRI) and cognitive performance data. They refer to these metrics as Alzheimer's Disease Pattern Similarity (AD-PS) scores. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), they calculated conditional probabilities modelled by large-scale regularised logistic regression. They conclude that AD-PS metrics could be a powerful tool in Alzheimer's Disease research to detect Alzheimer's Disease-like cognitive and anatomical effects and that this approach could be extended to other diseases such as Parkinson's disease.

In their 2011 paper, Ye *et al.* [17] claim that recent studies have demonstrated that imaging parameters based on brain scans are more consistent and more sensitive measures of AD diagnosis and progression than cognitive assessment. They report on the use of neuroimaging techniques including Structural Magnetic Resonance Imaging (sMRI), to measure specific structures such as the hippocampus, entorhinal cortex and amygdala and any abnormal volumetric changes related to Alzheimer's Disease. Another technique reported is the positron emission tomography (PET) scan, which uses different radioactive tracers to provide information on various physiological, biochemical and metabolic processes.

The Alzheimer's disease Neuroimaging initiative, described in [18] began in 2004 and had the overall objective of characterising clinical, genetic, imaging and biochemical biomarkers of the disease and identifying the relationships between them over the course of disease progression from normal cognition to MCI to dementia. It also established repositories of data and biological samples, both of which were to be freely accessible to the wider academic and research community. A possible use of these repositories as a basis for investigations using machine learning is to use them as training sets for the creation of classifiers such as SVMs (Support Vector Machines) or Decision trees.

Kehoe *et al.* [19] note that "structural MRI measures of the hippocampus and medial temporal lobe are still the most clinically validated biomarkers for Alzheimer's Disease, but newer techniques such as functional MRI and diffusion tensor imaging offer great scope in tracking changes in the brain, particularly in functional and structural connectivity, which may precede gray matter atrophy." This is quite an important statement and reflects the current viewpoint on neuroimaging based diagnosis of AD.

In 2015 the Informatics core at ADNI [20] published a review of the first decade of their data collection and dissemination. In the review they report that ADNI disseminates data to a continually growing number of investigators who have written hundreds of scientific papers based on ADNI data. The ADNI itself [11] reports that approximately 500 papers have been published as a direct result of ADNI to the end of 2013. Research using ADNI data crosses many scientific disciplines, geographic regions, and includes computer scientists interested in developing and testing machine learning and classification algorithms, neuroscientists interested in developing and testing models of disease

progression, radiologists, geneticists, and many others seeking to expand the boundaries of scientific knowledge.

PredictAD [21] is an EU funded project which aims to study imaging biomarkers (MRI, PET FDG and PET PIB), electrical brain activity measurement and blood based markers (proteomics and metabolomics) and develop methods for how to combine data from different biomarkers.

Several papers also report on the use of biomarkers/neuroimaging in the detection of Alzheimer's Disease including the following [22–29]. The ADNI database was utilised in many of these papers, mostly to train various machine learning classifiers

In the ADNI annual report for 2015 [2], section 2.2.8.4, reports on progress towards implementing the revised diagnostic criteria and evaluating biomakers [9]. They report that since the revised criteria were published in 2011, dozens of scientists have published results of studies implementing the revised criteria in research settings, examining the accuracy of biomarker tests in detecting and predicting Alzheimer's Disease and in distinguishing it from other forms of dementia. They conclude that although additional studies are needed before the revised criteria and guidelines are ready for widespread use in physicians' offices, preliminary evidence supporting the revised criteria and biomarker tests is growing.

Klopper *et al.* [30] reported on the application of machine learning techniques to neuroimaging-based diagnosis. The methods they studied promise fully automated, standard PC-based clinical decisions, unbiased by variable radiological expertise. They used Support Vector Machines (SVMs) to separate sporadic Alzheimer's Disease from normal ageing and from fronto-temporal lobar degeneration (FTLD). In their study, they compared the results to those obtained by radiologists. A binary diagnostic classification was made by six radiologists with different levels of experience on the same scans and information that had been previously analysed with a Support Vector Machine (SVM) based classifier, as we describe in a later chapter. SVMs correctly classified 95% (sensitivity-specificity: 95%-95%) of sporadic Alzheimer's disease and controls into their respective groups. Radiologists correctly classified 65–95% (median 89%; sensitivity-specificity: 88%-90%) of scans. SVMs correctly classified another set of sporadic Alzheimer's disease in 93% (sensitivity-specificity: 100/86) of cases, whereas radiologists ranged between 80% and 90% (median 83%; sensitivity-specificity: 80/85). SVMs were better

at separating patients with sporadic Alzheimer's Disease from those with FTLD (SVM 89%).

Although a great deal of research has been completed in the search for suitable biomarkers using these techniques, much work remains to be done. Sperling *et al.* [31] (Section 10) recommends the need for additional study, and in particular in the area of CSF arrays and PET/MRI analytic techniques. They mention significant challenges in implementing standardised biomarker cut-off values worldwide. They also entered a caveat concerning the research studies they used in preparing their recommendations. Specifically, in Section 8, they noted that although the studies provide compelling evidence that for the hypothesis that markers of A α and other specific factors might predict those individuals who are at a higher risk of progression to Alzheimer's Disease, there were several potential confounding issues in the majority of the studies that were not taken into account. Also, they note that the studies used individuals who self-selected for the research and consequently are not representative of an older population in general.

There are other factors that we must take into consideration regarding the biomarker and neuroimaging techniques that call into question their practicable usability:

- They are expensive and as a consequence it would be too costly to refer each patient for an MRI or PET scan and/or a spinal tap;
- As a result of the cost, the waiting lists for access to MRI and / PET scans are long and extensive;
- People, especially older people, might find the process of a neuroimaging scan, claustrophobic (see Figure2.2). They might also find, in the case of a spinal tap, the process intimidating and/or painful.

2.2 Other Techniques for Diagnosis of Alzheimer's Disease

Having reviewed the literature concerning the use of neuroimaging and other invasive techniques for diagnosis of Alzheimer's Disease, in this section we will review some other techniques found in recent literature. Laske *et al.* [32] investigated the need for additional non-invasive and/or cost-effective diagnosis tools. They stressed the points

FIGURE 2.2: Patient being loaded into an MRI scanner.

made above that current state-of-the-art diagnostic measures of Alzheimer's Disease are invasive (cerebrospinal fluid analysis), expensive (neuroimaging) and time-consuming (neuropsychological assessment) and thus have limited accessibility as frontline screening and diagnostic tools for Alzheimer's Disease on a large scale. Their paper analysed the number of available geriatrician, neurology and psychiatry physician providers, and also considered the number of MRI and PET imaging machines available in the USA and conclude there is not sufficient capacity available for any of these resources to provide comprehensive frontline screening tools for the at risk population. It mentions that many researchers have suggested that Alzheimer's Disease alone could bankrupt many medical systems if nothing is done immediately to develop inexpensive and/or non-invasive screening tools that do not require a specialist or specialised hardware.

The paper discusses other screening/diagnosis techniques under the two headings of neuropsychometric tests, and speech testing. We discuss the techniques under these headings below, with emphasis on the use of machine learning applications.

2.2.1 Neuropsychometric Tests

In their paper, Laske *et al.* [32] report on the use of standardised tests such as the logical memory subset from the Wechsler Memory scale, the Californian Verbal Learning Test, Free and Cued Selective Reminding Test and the Mini-Mental State examination. They compare the effectiveness of these standard tests for the detection of the various clinical stages of Alzheimer's Disease in two classes of potential patients; those that are

descendants of carriers of the PSEN1 E280A gene mutation and those that that are not, i.e those with potential sporadic Alzheimer's Disease. They reach some conclusions as to the optimal tests to use when screening these different classes of patients, but their paper does not investigate how Machine Learning (ML) methods could be applied in conjunction with the results of these or other standard tests.

In 1996, Datta *et al.* [33] ran a set of experiments using what were then the best available machine learning methods on the responses to standardised tests to refine the results of these tests in order to better screen normal brain aging from the earliest stages of Alzheimer's Disease. Their stated goal was to analyse machine learning methods to determine if they can improve the accuracy of dementia screening tools recommended by the *Agency for Health Care Policy Research*, (AHCPR).

Their experiments used the database maintained at the *Alzheimer's Disease Research Center* at the University Of California, Irvine. This contains the results of the initial visits of 578 possible patients and controls (either community volunteers or caregivers). Using the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition **(DSM-IV)**, the subjects' dementia status was classified as either normal, cognitively impaired but not demented, or demented.

They then used the responses and labelled classification stored in the database to generate datasets for the machine learning algorithms. The subjects' age, gender, job, and education along with the responses to the two tests were extracted from the database. These data constitute the attributes of the examples used by the machine learning algorithms. The standard tests used were the Functional Activities Questionnaire **(FAQ)** and the six-item *Blessed, Orientation, Memory and Concentration* exam **(BOMC)**. Both of these are recommended for screening by the **(AHCPR)**. These tests provide a method for assessing the functional and cognitive abilities of the subject.

Among the algorithms used were C4.5, C4.5 Rules and Naive Bayes (a description of these follow in the next chapter). The algorithms generate a classifier for the three dementia classes (normal, impaired but not demented, demented) from the training samples which are randomly selected from the 578 examples in the database.

The machine learning methods were applied in conjunction with the two tests and they report that their results show a 15-20% increase in classification accuracy over applying

either the cutoff criteria for **(FAQ)** (> 8 is demented), the **(BOMC)** (> 10 is demented) or their combined cutoff criteria in isolation. Combining the two tests **(FAQ & BOMC)**, results in a 60% improvement in classification accuracy. Their results show that Naive Bayes algorithm performed the best in classification accuracy using a larger training sample size, followed closely by the C4.5 and C4.5 rules algorithms. The C4.5 rules algorithm generates an easily understandable representation of the dementia status described with the attributes.

The researchers conclude that, unlike the **AHCP** criteria, machine learning methods can separate the cognitively impaired status from normal. This is important, as the high risk cognitively impaired can then be targeted for medical attention at an earlier time. Also, **(ML)** methods in conjunction with the results of the **FAQ** and **BOMC** can be applied to create simple statements to help classify the dementia status of patients.

The Datta *et al.* paper was published in 1996. However in a 2010 paper [34] Joshi *et al.*, noted that, although machine learning systems have been applied to a number of medical areas, dementia has not been one of them. It then puts forward the usage of various machine learning methods for the classification of the different stages of Alzheimer's Disease (mild, moderate, severe and Normal), similar to the earlier paper [33]. It references the earlier paper in it's literature survey and the experiments described in the paper are also similar. The methodology follows the earlier paper [33] viz-a-viz :-

a) A Database maintained by the National Institute of Aging in the USA contains the responses from the initial visits of 496 subjects, seen as controls or patients. Using the **DSM-IV** criteria, the subjects were classified as normal, cognitively impaired or dementia of the Alzheimer's Disease type. A process of attribute selection preceded the generation of training and test sets for the machine learning algorithms. This was done using a gain-ratio attribute evaluation scheme with ranker search method for selecting the attributes. In all, 35 attributes were broadly classified under five main observations namely age, neuropsychiatry assessments, mental status examination, laboratory investigations and physical exmainations.

b) After the attributes were selected, different models were simulated using various machine learning methods such as decision tree (C4.5), the bagging method, Neural Networks, Multi-layer Perceptrons **(MLPs)**, CANFIS [34] and Genetic Algorithms.

The test set of 225 subjects were classified as normal, mild cognitive impairment, moderate impairment and severe. One case was mistakenly classified as moderate when it should have been classified as mild The optimal classification accuracy was found to be using **CANFIS**, at 99.5%, closely followed by **MLPs** at 98.99% and **C4.5** at 98.97%. The performance of C4.5 is interesting when compared to it's performance in the earlier results reported by Datta above.

2.2.2 Speech Testing

Alzheimer's Disease produces a variety of communication problems in spoken language, including aphasia (difficulty speaking and understanding) and anomia (difficulty recognizing and naming things) [9].

Jarrold *et al.* [35] in 2015 studied the use of computational analysis of language as a diagnostic for brain-based disorders. They propose that word choice and other linguistic markers are heavily affected by these disorders.

They present a machine learning-based methodology for identifying and testing language measures that serve as markers for brain-based disorders. They then evaluated the application of the methodology to the three disorders: pre-symptomatic Alzheimers Disease (pre-AD), cognitive impairment and depression. They claim that the methodology independently discovers a relationship previous reported in the literature and that it produces accurate diagnostic models. The method allows researchers to classify patients according to patterns in speech and language production. The process they implemented required that audio recordings of structured interviews with the patient were transcribed into raw text. Each interview was annotated with the diagnosed label for the patient. They then applied various lexical feature extraction tools to the text to produce a Lexical Feature vector. The vectors were fed into implementations of three machine learning algorithms: logistic regression, J48 (an implementation in Weka of the C4.5 algorithm), and a multi-layered perceptron.The software used was the open source Weka machine learning toolkit.

The paper has a separate section reporting on experiments concerning Alzheimer's Disease and cognitive impairment in which they focus on results they obtained that indicate

that induced machine learning models trained by the linguistic features can detect current cognitive impairment and predict future onset of Alzheimer's Disease.

Citing [36] (1996), they refer to the findings by Snowdon *et al.* (known as the Nuns study) that a language characteristic known as low idea density in the autobiographical writings of American Nuns in their 20's was a strong predictor of Alzheimer's Disease at the time of their death more than 50 years later. This information provided the basis for their aim to predict preclinical Alzheimer's Disease from language analysis. They selected a sub-sample of 22 subjects who were cognitively normal at the time of the 1988 interviews but eventually died with the cause of death listed as clinically verified Alzheimer's Disease. They selected controls with an aged-matched cognitively normal sub-sample of 23 men never diagnosed with dementia.

The lexical analysis tool **CPIDR** [37], used in the feature extraction phase, is used to measure idea density in the text. They found, as hypothesised, that idea density as measured by CPIDR was significantly less in the Pre-Alzheimer's Disease group than in the matched controls.After applying machine learning models trained by the features generated by the lexical analysis of the speech transcripts,they claim they were able to predict which individuals went on to develop Alzheimer's Disease, with an accuracy of 73%. They state that, in their opinion, this was their most significant result from their experiments.

Lopez-de-Ipena *et al.* (2013) [38] report on the use of automatic speech analysis techniques and demonstrate how it can be performed (after proper training) by anyone in the patient's habitual environment without altering or blocking the patient's abilities. The technique only requires verbal tests and interviews with the patient. It can also help to estimate the severity of Alzheimer's Disease in the patient, as the specific communication problems a patient encounters depend on the stage of the disease, e.g, a common symptom at the mild cognitive Impairment (MCI) stage of the disease is that the patient has trouble finding the right word during spontaneous speech.This often remains undetected. Another area that is affected early in the disease stage is emotional responsiveness. The deterioration of spoken language immediately affects the patient's ability to interact naturally within his or her social environment, and is usually also accompanied by alterations in emotional responses. Often one observes social and behavioral changes in the early stages of the disease. Both of these changes appear early

in the progress of Alzheimer's Disease and both can be measured according to Horley [39].

Other studies of the use of speech analysis include [37], published in 2008. This study used machine learning methods known as deep-belief networks (DBNs), which are a form of automatic neural network. They also used logistic regression and claim to obtain 70.9% and 77.6% accuracy on average respectively using these algorithms

A paper by Baldas *et al.* [40] in 2010 adapts a similar approach to that of Jarrold *et al.* [35], in that voice recordings are transcribed into raw text for input into lexical analysis tools. Various stylometric measures are generated using methods from the Linguistics field. Thomas *et al.* [41] in 2005 has shown that the stylometric attributes have sufficient discriminating power in distinguishing the language models of Alzheimer's Disease patients and control subjects. The study presents a method for constant monitoring of a subject's speech that can analyze the lexical data and decide on whether or not his cognitive status may be deteriorating due to Alzheimer's Disease.

The Thomas*et al.* paper described a detailed statistical analysis of the lexical features in the spontaneous speech of older adults with Alzheimer's Disease.It also used several machine learning and natural language processing techniques in rating Alzheimer's Disease, and the researchers implemented their own classification algorithm, which they called Ordinal CNG. They reported positive results in that several standard classification algorithms could be used to produce high classification accuracies.

2.3 Summary

In this chapter we have reviewed different approaches to diagnosis of Alzheimer's Disease. These included techniques based on biomarkers and based on neuroimaging, both of which are expensive and around which there are questions of whether there actually is a practical case to be made for their use. We also looked at the various neuropsychometric tests including analysis of speech and dialogue.

For all of these diagnosis options we can conclude that they are complex and the need for expertise makes them expensive, as we try to support sound decusion-making. In

the next chapter we hive an overview of machine learning techniques and this will then allow us to set out the research questions behind our work, which we do in Chapter 3.

Chapter 3

Machine Learning Overview and Tools

3.1 Introduction to Machine Learning

Machine Learning is a branch of Artificial Intelligence that has become very popular, and useful, in the last 10 years. One definition of Machine Learning is that it is the semi-automated extraction of knowledge from data. Broadly speaking, machine learning (ML) deals with the question of how to build computer programs that learn from data and, as a result, can generate programs that generalise from that data in the form of a program that reflects concepts implicit in the underlying data. In effect, with machine learning we have programs using data to create new programs. This is in contrast to the traditional way that programs have been generated by human programmers in which they encode the rules that the computer follows in a programming language in order to produce a solution to a specified problem.

Traditional or conventional writing of programs for a computer can be summarized as automating the procedures to be performed on input data in order to create output artifacts. Almost always, they are linear, procedural and logical. A traditional program is written in a programming language to some specification, and it has properties like:

- You know or can control the inputs to the program;
- You can specify how the program will achieve its goal;

- You can map out what decisions the program will make and under what conditions it makes them;
- You can test your program and be confident that, because the inputs and outputs are known and all conditions have been exercised, the program will achieve its goal.

The top half of Figure 3.1 below illustrates this process

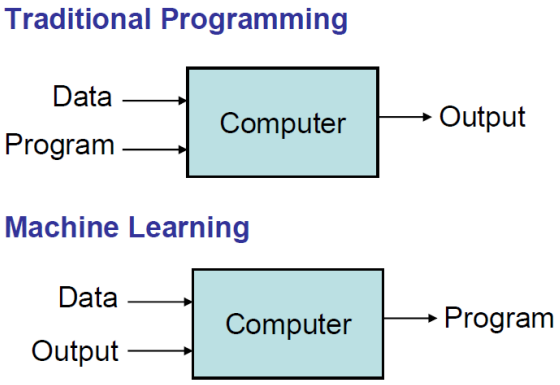

FIGURE 3.1: Traditional programming vs. machine learning

Traditional programming works on the premise that, as long as we can define what a program needs to do, we are confident we can define how a program can achieve that goal. This is not always the case as sometimes, however, there are problems that you can represent in a computer that you cannot write a traditional program to solve. Such problems resist a procedural and logical solution. They have properties such as:

- The scope of all possible inputs is not known beforehand;
- You cannot *specify* how to achieve the goal of the program, only what that goal is;
- You cannot map out all the decisions the program will need to make to achieve its goal;
- You can collect only sample input data but not all possible input data for the program

Problems like this resist traditional programmed solutions because manually specifying a solution would require a disproportionate amount of resources. Furthermore, when new inputs arise, the rules may change, thereby necessitating changes to the program.

In such cases as these, machine learning might be the optimum approach to use in deriving a solution to the way the problem is represented on the computer, and that is what we focus on in this chapter.

There are two broad classes of machine learning techniques ... *supervised learning* and *unsupervised learning*.

Supervised learning takes a set of feature/label pairs, called the training set. From this training set the system induces a generalised model of the relationship between the set of descriptive features and the target features in the form of a program that contains a set of rules. The objective is to use the output program produced to predict the label for a previously unseen, unlabelled input set of features, i.e. to predict the outcome for some "new" data. The features correspond to the input named "data".

In the second half of Figure 3.1, the input named "output" are the labels. When we run a machine learning algorithm on data which has been collected, the algorithm attempts to create a program or a model that knows how to solve the problem. This is the output "program" in the second half of Figure 3.1. In this case the box named "computer" would be one of a collection of algorithms designed to solve machine learning problems. Figure 3.2 below shows an example of supervised learning in action. The first three figures on the left comprise the training set and are labelled as "chairs".

These examples are used by the system to "learn" the characteristics of chairs so that, when presented with the fourth figure on the right, it classifies it as a chair also. This is an example of a supervised classification problem.

FIGURE 3.2: Generalising from examples of chairs

Figure 3.3 below shows this process as a as a workflow. The training data with the label of "chairs" in the top left of the figure represents the chair figures, and the combination of the data and the labels are used to generate the model. The model is then used to predict the label for the fourth, unknown or unlabelled figure.

The bottom half of Figure 3.3 actually illustrates a common way to evaluate the accuracy of the generated model. Data with known labels, which have not been included in the training set, are classified by the generated model and the results are compared to the known labels. This dataset is called the test set. The accuracy of the predictive model can then be calculated as the proportion of the correct predictions the model labeled out of the total number of instances in the test set

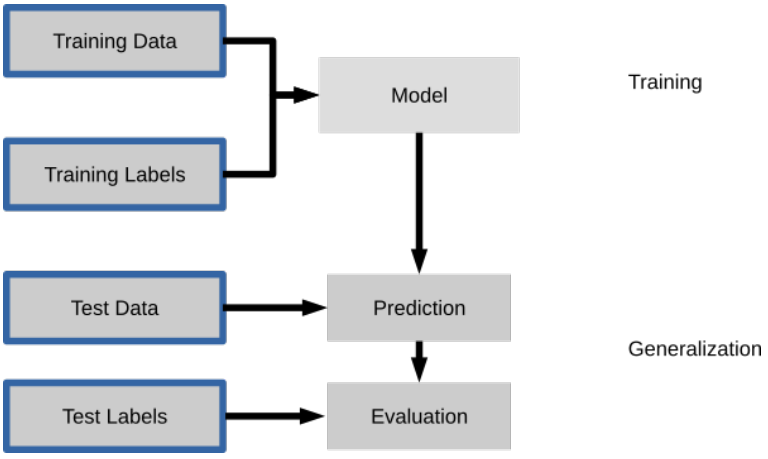

FIGURE 3.3: Supervised machine learning workflow

Unsupervised learning is the second form of machine learning and also takes a dataset of descriptive features, but without labels, as a training set. The goal now is to create a model that finds some hidden structure in the dataset, such as natural clusters.

Machine learning algorithms are tools to automatically make decisions from data in order to achieve some over-arching goal or requirement. The promise of machine learning is that it can solve complex problems automatically, faster and more accurately than a manually specified solution, and at a larger scale. Over the past few decades, many machine learning algorithms have been developed by researchers, and new ones continue to emerge and old ones modified. Figure 3.4 below, taken from [1], attempts to group some common machine learning algorithms by their similarity in form or function. This covers some of the recently developed algorithsm like the variations on Deep Learning, as well as some of the older, classic algorithms like CART and C4.5 Decision Trees. The graphic also covers many different ways in which machine learning can be applied, from clustering, to developing decision trees for human perusal, to rule-based systems which are not meant to be viewed by people, as well as techniques like dimensionality reduction and regression.

From all this variations however, in our research we will be concerned solely with algorithms used in solving supervised classification problems given that we have groundtruth (answers) which can act as the supervision, and our target is to perform a classification process.

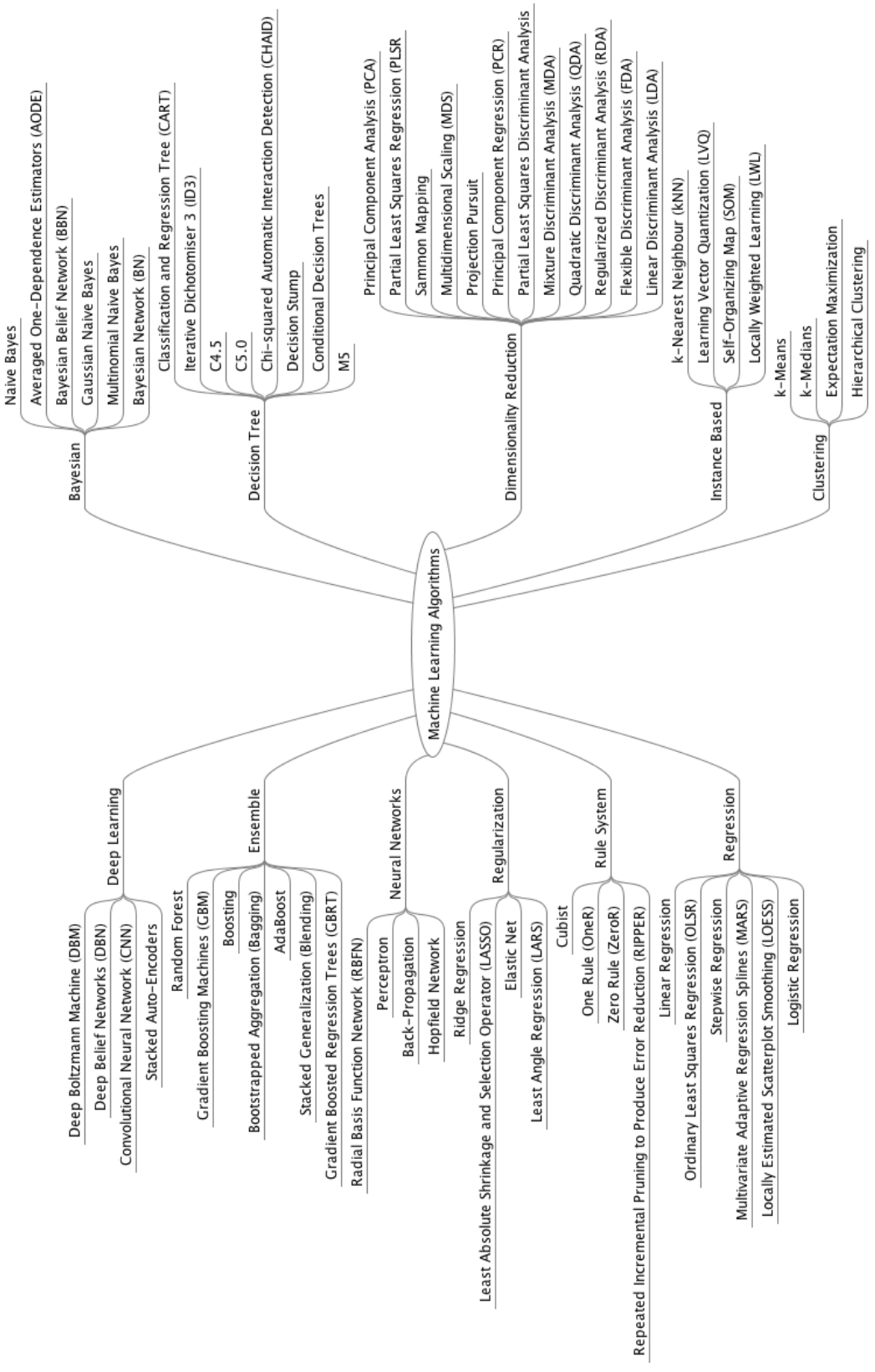

FIGURE 3.4: Machine learning algorithms grouped by type (from [1])

In the work reported here we are attempting to find an answer to a specific research question namely *"Can machine learning techniques be used to screen subjects for the onset of Alzheimer's Disease at an early stage of the disease and what is the data we need to do this ?"*. We propose to address answering the question as a standard machine learning task, which will have a significantly greater chance of a successful completion if we use a standard procedure to manage the project through the project lifecycle.

Therefore, the project of exploring the research question will be organised using an industry standard process — the Cross Industry Standard Process for Data Mining (CRISP-DM). While the name refers to data mining, the process can also be applied to a machine learning project. It is designed so that a project can be broken down into manageable phases and consists of checklists, guidelines, tasks, and objectives for every stage of the project. It is expected that the project outputs, when the project is run using CRISP-DM, will consist of deployable, reproducible and documented research.

Figure 3.5 below show the process of running a project using CRISP-DM in diagrammatic form, and is fairly self-explanatory. The remainder of this chapter, and subsequent chapters, will describe the first phases of the CRISP process, the process of understanding, getting and preparing the data required for the research.

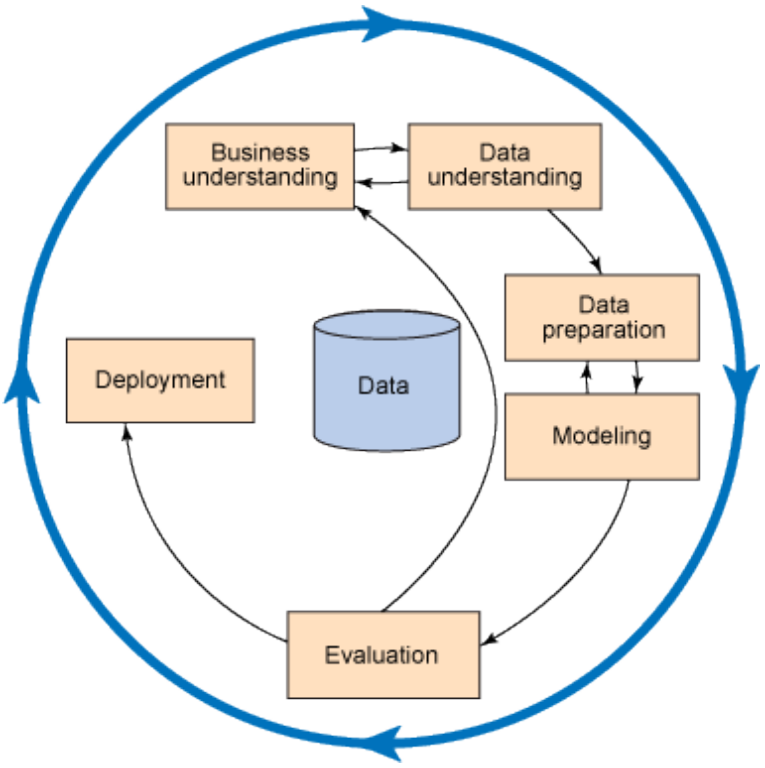

FIGURE 3.5: The CRISP data mining process

3.2 Machine Learning Tools and Scikit-Learn

There are many different software tools available to build machine learning models and to apply these models to new, unseen data. There are also a large number of well defined machine learning algorithms available, some of which are in Figure 3.4. These tools typically contain libraries implementing some of the most popular machine learning algorithms and can be categorised as follows :

- pre-built application-based solutions or, alternatively,
- programming language based with specialised machine learning libraries.

The former group includes well-known frameworks such as WEKA, Knime Analytics Platform, RapidMiner Studio and SAS Enterprise Miner. WEKA is open source and very popular among researchers and developers. It is used and referred to in many of the research papers found in the literature. Initially we used this package to build classifiers and found that, although it was easy to use and get things done, the problem was that the framework was basically designed as a black box and it was difficult to understand and explain the results we were getting. Using WEKA it was possible to train and evaluate a model very quickly. However, the programming language approach to developing and implementing models is more flexible and gave us better control of the parameters to the algorithms. It also allowed us to have a better understanding of the output models produced. The programming options we looked at included Python with the popular SciKit-Learn library[1]. This again is open source and regularly tops polls such as those on the KDnuggets website[2] relating to usage of machine learning languages. An alternative programming language we considered was the Statistical language R, but Python was our preferred choice, primarily because of the SciKit-learn library extensions and the documentation available with it.

Since its release in 2007, scikit-learn has become one of the most popular open source machine learning libraries. Scikit-learn (also called sklearn) provides algorithms for machine learning tasks including classification, regression, dimensionality reduction and

[1] http://scikit-learn.org/stable/
[2] http://www.kdnuggets.com/

clustering. It also provides utilities for extracting features, processing data and evaluating models. It provides in-built code for many of the algorithms contained in Figure 3.4 above. The documentation for scikit-learn is comprehensive, popular and well maintained. Sklearn is built on mature Python Libraries such as NumPy, SciPy, and matplotlib. It has a very active development community with regular update releases of the library. As of 2015, scikit-learn is still under active development and is sponsored by INRIA, Telecom ParisTech and occasionally Google (through the Google Summer of Code).

Another popular language used for machine learning is R. We choose scikit-learn as the tool to use for our experiments because:

- We already have some familiarity and exposure to Python, and thus have a smaller learning curve;
- The excellent documentation and tutorials available;
- The number of classic machine learning algorithms that come with it, and the consistent patterns (API) for using the different models e.g each model can be used with the same basic commands for setting up the data, training the model and using the model for prediction. This makes it easier to try a range of machine learning algorithms on the same data;
- The machine learning algorithms included with sklearn have tunable parameters known as hyperparameters that effect the performance of the model. These usually have sensible default values, so that we can run them without needing a detailed knowledge or understanding of their semantics;
- The IPython notebook, which is an interactive computational environment for Python, in which a user can combine code execution, rich text, mathematics and plots in a web page. This functionality allows us to provide the notebooks we used to run our experiments almost as an audit and in a presentable and easily understood way that allows for reproducible research. The notebooks are explained in the following chapters.

3.3 Making Predictions from Memory Works Clinic Data

We used data from the Memory Works clinic described earlier in Section 1.3, as a basis for training a supervised classifier applying the scikit-learn environment described above. This generated an IPython notebook which contains the Python code developed to make predictions with the data. The objectives were to establish the level of accuracy of prediction that we would obtain with the existing data and from that to establish if the Memory Works clinic database contained a sufficient number of instances (subjects, or rows) and features (answers to questions, or columns) in order to train a machine learning algorithm such that it could perform a reasonably accurate predictions on previously unseen data.

The accepted way in which to measure such accuracy of prediction on unseen data, using an existing dataset, is to perform an *n-fold cross-validation* (sometimes referred to as k-fold). Cross-validation is a fairly standard statistical technique used to measure classifier accuracy. It involves randomly dividing a dataset (rows or subjects in our case) into n equally-sized subsets and then taking $n-1$ of these subsets to train a classifier, which is then run on the n^{th} subset and the accuracy of the classification on this n^{th} subset is measured. We apply this process n times, leaving out each of the n subsets in turn from the training data, and combining the prediction accuracy figures for each of the n subsets to get an overall classifier prediction accuracy.

So, for example, in a 10-fold cross-validation from a dataset of 120 instances, we leave out instances 1 ... 12, train on instances 13 ... 120, and test performance on instances 1 ... 12. We then leave out instances 13 ... 24, train on instances 1 ... 12 + 25 ... 120, and test on instances 13 ... 24, and so on, repeating the process 10 times in total.

The reader should note that although this iPython notebook looks fairly short and straightforward, what we are seeing is the result of many iterations using scikit-learn, and the iPython notebook is in fact replaying the final version of the work. The notebook is generated as a series of numbered cells (numbered like this ... [4]) and we present the numbered cells in sequence, starting with cell [2], with each bounded by a box and in a reduced font size. Note that the code was run on a Linux system so the command line instructions are in Unix.

```
In [2]: ! ls data
        activity.csv impression.csv main table.csv outcomes.csv
        drug history.csv lifestyleactivity.csv med history.csv
        psych history.csv
```

The first command we execute is to list the CSV tables which were generated from the database schema described earlier in Figure 1.1, in Chapter 1. We then import the necessary Panda module and list the comma separated files (csv) that have been created from the MySQL database storing the Memory Works clinic data, all of which is done in cell [4] below, and we also output the number of rows, columns and unique subjects.

```
In [4]: main_table = pd.read_csv('data/main_table.csv')
        print("orginal number of rows", len(main_table))
        # Drop duplicate columns (generated by join)
        duplicate_cols = [i for i in main_table.columns if "." in i]
        main_table.drop(duplicate_cols, axis=1, inplace=True)

        # Ensure patientID not 0
        main_table = main_table[main_table['patientID'] != 0]

        # Ensure age older than 20
        main_table = main_table[main_table['age'] > 20]

        # The following columns cannot be null
        not_null = ["analysisID", "FormID", "patientID"]
        main_table = main_table.dropna(subset=not_null, how="any")

        print("Filtered Number of rows: ", len(main_table))
        print("Number of columns: ", len(main_table.columns))
        print("Unique subjects: ", len(main_table['patientID'].unique()))

orginal number of rows 190
Filtered Number of rows:  123
Number of columns:  841
Unique subjects:  123
```

Cell [4] above reads the main table and takes some steps to ensure integrity of the data. It filters the data and, after dropping duplicate entries and invalid data, it results in a Pandas dataset with 123 unique rows representing subjects with 841 columns for features.

```
In [6]: id_columns = [i for i in main_table.columns if 'ID' in i]
                      print("ID columns: ")
                      print(id_columns)

             ID columns:
               ['detailsID', 'form FormID', 'FormID', 'patient patientID',
               'patientID', 'clinic ClinicID', 'clinician cl
```

Above in cell [6] we list column names from resulting database

```
In [7]: # Read other datasets
# Note: each of these will eventually become a single, or multiple
features in the main table
                      activity = pd.read_csv('data/activity.csv')
                      drug_history = pd.read_csv('data/drug_history.csv')
                      impression = pd.read_csv('data/impression.csv')
                      lifestyle_activity = pd.read_csv('data/lifestyleactivity.csv')
                      med_history = pd.read_csv('data/med_history.csv')
                      psych_history = pd.read_csv('data/psych_history.csv')
                      outcomes = pd.read_csv('data/outcomes.csv')
```

In cell [7] above we read the other tables exported from the database

```
In [8]: features = main_table[:];;'FormID']]

# age_feature
features['age'] = main_table['age']

# family_present
features['family_present'] = main_table[['family_present']]

# is_male
features['is_male'] = (main_table['gender']=='male')*1

#retired
features['is_retired'] = main_table['occupation'].apply(lambda x: 1
                                if (type(x)==str) and 'retir

# kin response
features['kin_concerned'] = (main_table.kin_response=='concerned')*1
features['kin_not_noticed'] = (main_table.kin_response=='not_noticed')*1

# Checks
# For every column with "check in its name, append to features"

check_columns = [i for i in main_table.columns if 'check' in i.lower()
                                            and 'collat' not in i.low
        for i in check_columns:
            # If the feature is too sparse (less than 15\% positive results),
            Then:
            do not include feature
            if main_table[i].dtype != 'O' and main_table[i].fillna(0).mean() > 0.15:
                features[i] = main_table[i].fillna(0)
```

In the above we start to extract features and target variables from the data. Cell [8] reads in the other csv files and cell [9], below, extracts features and target variables from the data. It creates a Pandas dataframe called 'features'.

```
In [9]: print("feature avgs. ")
        features.describe().loc['mean', :]

 feature avgs.

Out[9]:     FormID              97.626016
            age                 65.333333
            family_present       0.024390
            is_male              0.308943
            is_retired           0.455285
            kin_concerned        0.430894
            kin_not_noticed      0.219512
            anxiety_check        0.463415
            comments_check       0.373984
            decisions_check      0.414634
            faces_check          0.243902
            follow_conv_check    0.528455
            losing_things_check  0.715447
            names_check          0.747967
            rec_events_check     0.439024
            right_words_check    0.154472
            falling_check        0.162602
            headaches_check      0.219512
            bereavement_check    0.398374
            Name: mean, dtype: float64
```

In the above cell [9] we print the average value for each column. For example, the average age is 65.3. For Boolean features the mean represents what percentage of respondents answered yes. So, for example, is_male is 0.31. So 31% of respondents are male. Lets now look at some of the actual data.

```
In [10]: # Sample of features
         features.iloc[:5, :]

Out[10]:    FormID  age family_present  is_male is_retired  kin_concerned \
        17  158     68  0               1       0           0
        19   86     64  0               0       1           0
        20   85     70  0               0       1           1
        21   99     69  0               0       0           0
        22   96     80  0               0       1           1

            kin_not_noticed anxiety_check   comments_check  decisions_check \
        17  1               0               0               0
        19  0               1               1               1
        20  0               1               1               1
        21  0               1               0               0
        22  0               0               0               1

            faces_chk    follow_conv_chk   losing_things_chk names_chk \
        17        0                   1                   1           1
        19        0                   0                   1           1
        20        0                   0                   0           1
        21        1                   0                   1           1
        22        0                   1                   1           0

            rec_events_chk  right_words_chk falling_chk headaches_chk \
        17         0            1               0           0
        19         0            0               0           0
        20         1            0               0           0
        21         1            0               0           0
        22         1            0               1           0

            bereavement_check
        17     0
        19     0
        20     0
        21     0
        22     1
```

In the above cell [10] we look at a sample of features

```
In [11]: # Append relevant ID's
         # Remember to remove so we don't accidentally perform machine learning on index
         features['analysisID'] = main_table.analysisID
```

In the above cell [11] we append relevant foreign key columns to features. These will be useful when extracting features from other databases and will help the processing to run faster.

```
In [15]: outcomes.iloc[:5]
Out[15]:    outcomeID   outcome         outcome_notes       analysis_analysisID
        0   1           gp letter       GP letter notes     1
        1   2           counselling     NaN                 2
        2   3           counselling     NaN                 3
        3   4           counselling     NaN                 4
        4   5           counselling     NaN                 5
```

Cell [15] shows that each visit can result in multiple outcomes, which presents another challenge to our classification since the output can have multiple labels or outcomes.

```
In [16]: num_outcomes = outcomes.analysis_analysisID.value_counts().value_counts()

         for k, v in num_outcomes.iteritems():
             print("{} outcome(s) prescribed {} times".format(k, v))

         print("total: ", len(outcomes))

1 outcome(s) prescibed 68 times
2 outcome(s) prescibed 53 times
3 outcome(s) prescibed 12 times
4 outcome(s) prescibed 1 times
total:  214
```

In cell [16] we can see that we cannot simply ignore situations where multiple outcomes are prescribed. Hence we must treat this not as a 1-vs-all classification problem, but as a multi-label classification.

```
In [17]: # convert to multi-label dummies

         outcome_dummies = pd.get_dummies(outcomes.outcome)
         outcome_dummies['analysisID'] = outcomes.analysis_analysisID
         combined_outcomes = outcome_dummies.groupby('analysisID').
                                        apply(lambda x: x.sum())
         combined_outcomes = combined_outcomes.drop('analysisID', axis=1)

         print("Total num outcomes:", len(combined_outcomes))

         combined_features = features.merge(combined_outcomes, how='inner',
                                            left_on='FormID', right_inde
         print("Amount of subjects diagnosed: ", len(combined_features))

Total num outcomes: 134
Amount of subjects diagnosed:  93
```

Our way to approach the multi-classification situation is that we will need to split the answers into dummies, and then combine the answers for each subject, which is done above in cell [17].

```
In [18]: input_features = [i for i in features.columns if "ID" not in i ]
         prediction_variables = combined_outcomes.columns

         X = combined_features[input_features]
         y = combined_features[prediction_variables]
```

Finally, having set up the input data and the desired outputs, we generate the feature vector (X) and the target variables (y). So what we have done here is that after some further data manipulation and analysis in cells [10.. 17], we create the input_features and prediction variables in cell [18]. These are stored in the customary X & y variables for the sklearn algorithms.

```
In [19]: X.iloc[:5, :6]

Out[19]:
        age family_present  is_male     is_retired  kin_concerned   kin_not_noticed
    19  64  0               0           1           0               0
    20  70  0               0           1           1               0
    21  69  0               0           0           0               0
    22  80  0               0           1           1               0
    24  77  0               0           1           1               0
```

Above in cell [19] is a sample of the input matrix (feature vectors)

```
In [20]: y.iloc[:5]

Out[20]:
            coping_strategies   counselling gp_letter   has_diagnosis  leaflets \
    19      1                   0           0           0               0
    20      1                   0           1           0               0
    21      1                   1           0           0               0
    22      1                   0           1           0               0
    24      1                   0           0           0               0

            unknown
    19        0
    20        1
    21        0
    22        0
    24        0
```

Above in cell [20] is a sample of the target variables and their values, i.e. the known groundtruth.

Having explained the Memory Works clinic dataset and its pre-processing and loading into the system, we can now move on to some machine learning and cross-validation

```
In [21]: from sklearn.multiclass import OneVsRestClassifier
         from sklearn.svm import LinearSVC
         from sklearn.cross_validation import cross_val_score
         from sklearn.metrics import coverage_error
```

Th first step in the machine learning in cell [21] is to examine the target variables.

```
In [22]: y.describe().loc['mean']

Out[22]:    coping_strategies   0.913978
            counselling         0.204301
            gp_letter           0.397849
            has_diagnosis       0.032258
            leaflets            0.000000
            unknown             0.107527

In [23]: y = y.drop('leaflets', axis=1)
```

We can see from the values for the output or target variables that none of the subjects in the database were actually given leaflets as an outcome, so we will drop this as a target variable. Of the rest of the outcomes there are some important points to consider:

- **Has_diagnosis:** It is unclear what this means since only 3% of subjects were given this as an outcome, so the results of a predictor built on this will probably be poor;
- **Unknown:** is also unclear what this means since only 10% of subjects were given this, so the results of the predictor will also be unclear;
- **Coping_strategies:** Most subjects were given coping strategies (91%)

We will use Support Vector Machines to classify every target variable seperately. We could also treat this directly as a label classification problem using a one-vs-all classifier, but these are more difficult to score, and to rank their factors. We use the Receiver

Operating Characteristic Area Under Curve (ROC AUC) as an evaluation measure because it works better than other measures when the classes are biased, such as for the "has_diagnosis" and "coping_strategies" targets above. Scores above 0.5 are good

```
In [24]: from pylab import rcParams
         rcParams['figure.figsize'] = 15, 8

In [25]: svm = LinearSVC(C=10)

         data = []
         for i in y.columns:
             data += [cross_val_score(svm, X, y[i], cv=3, scoring='roc_auc')]
         pd.DataFrame(data, index=y.columns).T.boxplot()
pass
```

The above process in cells [24] and [25] uses a 3-fold cross validation and involves a machine learning process. To show the results from this we generate boxplots for the accuracy of predicting each possible outcome and this shows how well an SVM can predict each outcome, and is shown in Figure 3.6.

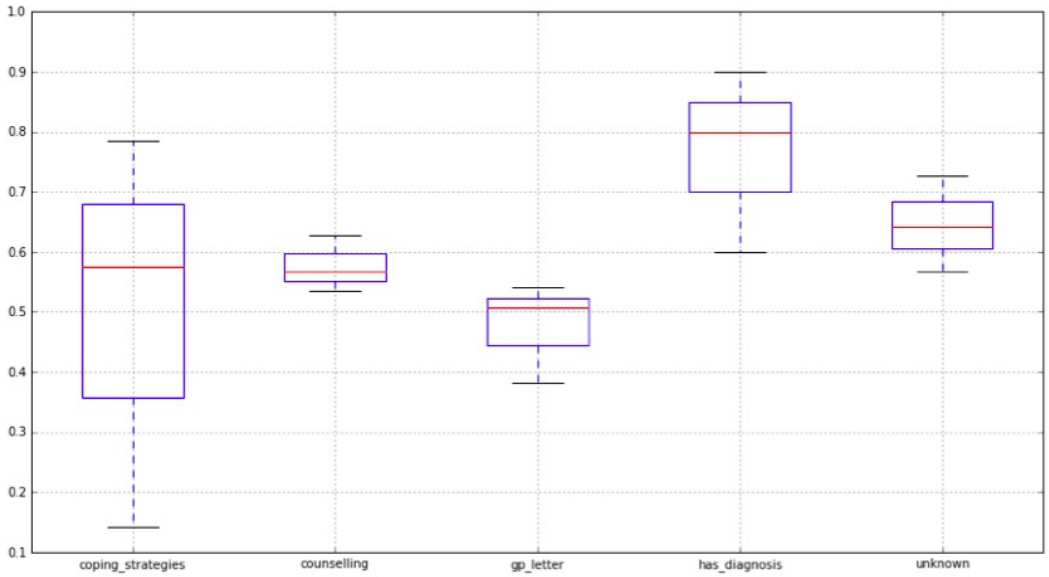

FIGURE 3.6: Boxplot of accuracies for predicting outcomes

From this Figure we can see that each of the boxplots represents the results of cross validation when predicting for one the the five target outcomes. The two right hand side targets, "has_diagnosis" and "unknown" look promising, but as only 3% of the subjects were labelled with the former and 10% of the letter ,there results are misleading.The first target, "coping strategies", was the label given to 91% of the subjects, scores well

over 0.5, but has a lot of variance, mostly at the lower end of the scoring scale.‘ The third boxplot, for the target “gp_letter”, is mostly below the 0.5 score. The second boxplot, “counseling” scores well over the 0.5 mark, but this label only accounts for 20% of the subjects.

Where this preliminary analysis leaves us is that we now need to examine the relative importance of the input features since not all features have equal impact on the classification outcome. This means weighing the importance of features for each outcome. So to do that here we use an ExtraTrees Classifier to fit the data, which we import in cell [27]. The Extra Tree’s classifier can handle multiple target variables, and can also rank the importance of each feature Below is the sorted list of most important features

```
In [27]: from sklearn.ensemble import ExtraTreesClassifier

In [28]: forest = ExtraTreesClassifier(n_estimators=250,
                                       random_state=0)
         forest.fit(X, y)
         importances = forest.feature_importances_
         importances = pd.Series(importances, index=X.columns)
         importances.sort(ascending=False)
         importances

Out[28]:    age                     0.141391
            anxiety_check           0.067509
            follow_conv_check       0.066740
            decisions_check         0.066448
            rec_events_check        0.061770
            is_retired              0.060288
            losing_things_check     0.059062
            headaches_check         0.057325
            is_male                 0.056209
            comments_check          0.055302
            kin_concerned           0.053768
            faces_check             0.051935
            kin_not_noticed         0.048184
            bereavement_check       0.047324
            names_check             0.042611
            falling_check           0.041865
            right_words_check       0.012760
            family_present          0.009510
```

Above in cell [28] is the is the sorted list of most important features in the classification of the Memory Works clinic data using a 3-fold cross-validation. Note that this is a multi-label classification. The relative feature importances are shown in the graph in Figure 3.7

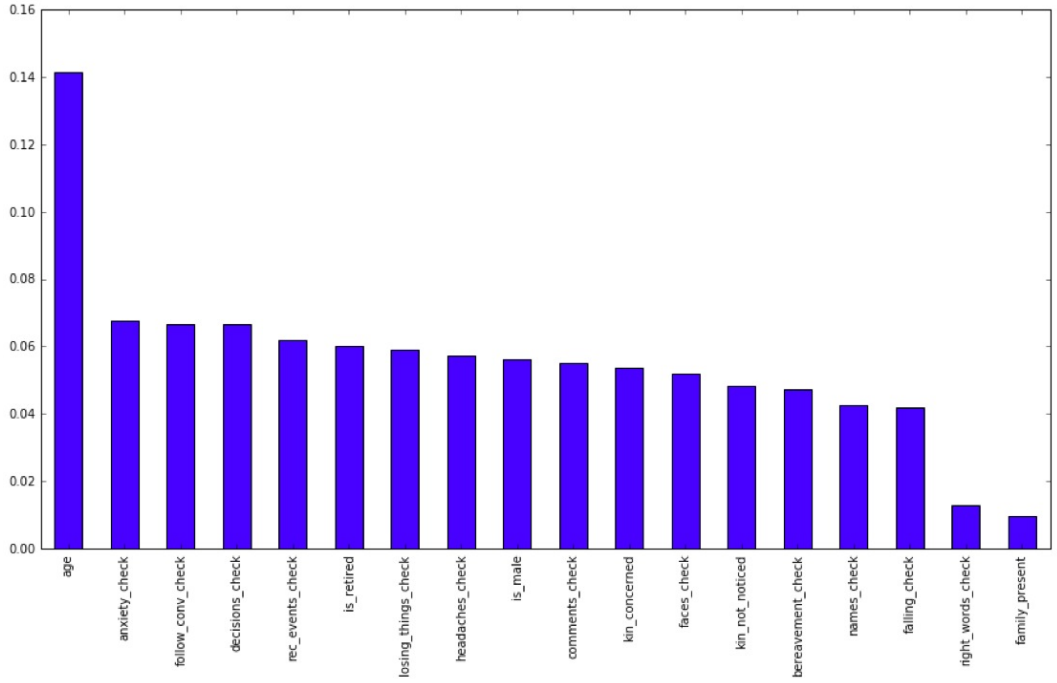

FIGURE 3.7: Relative importance of features in Memory Works clinic dataset

The analysis of the features used, which is shown in the graph of of Figure 3.7 shows that "age" is the most important feature, with all the other features contributing far less in the prediction. From this we conclude that in order to train an algorithm to generate more accurate predictions, we need a greater number of training instances with a greater range of meaningful feature values and perhaps even more features.

The task we are now faced with is to quantify both of these characteristics, and it is this task that leads us to another dataset, the NACC dataset, which we will use to establish the optimal quantities for number of instances and the most salient features to use in training the models. In the next Chapter we introduce the NACC dataset, but before that we are now in a position to introduce our research question , which is broken down into 4 sub questions in the next section.

3.4 Research Questions and Research sub questions

Here we introduce the Main Research Question and the 4 sub questions arising from this Question, which are re-examined in the concluding chapter ... refer to start of chapter 6 where they are revisited

Having carried out an initial analysis of the clinical data from the Memory Works clinic and examined how the data may be used to automatically mine patterns and build a system for processing new data we are now in a position to state our main research question, which is as follows:

> "Using clinical data gathered from subjects being assessed for possible dementia, it is possible to use machine learning to learn patterns which can be applied to clinical assessment of new subjects, and to predict a likely diagnosis".

This quite a bold statement, and following from that Question, there are a number of sub research questions that need to be addressed before we can arrive at an answer to the main question. Those research questions are as follows:

1. **RQ1:** What is the accuracy with which a subject's diagnosis of dementia can be predicted based on a set of clinical data for that subject, compared to patterns mined from a dataset of previous diagnoses ?
2. **RQ2:** Clinical assessment data for what number of subjects, is required in order to automatically mine patterns and make a reliable diagnosis for a new, as yet unseen, subject ?
3. **RQ3:** In carrying out a prediction of diagnosis outcome for a subject, which clinical features are more, or less, important ?
4. **RQ4:** What does the tradeoff between accuracy of diagnosis prediction vs. number of subjects in the training dataset look like and is there a "sweet spot" in that tradeoff ?

While some aspects of these research questions may appear as typical of any machine learning application, the nature of the data we are dealing with, the fact that it is noisy and sometimes inaccurate, that it varies, all this means that working with it presents some unique challenges, and of course as a societal challenge, any kind of assistance to diagnosis of dementia is of great importance.

We shall return to each of the research questions in the remaining chapters of the thesis and we return to the Main research Question in the final Chapter.

3.5 Summary

In this chapter we have presented an introduction to machine learning, covering the difference between traditional structured, algorithmically-based programming, and machine learning based on a data-driven approach to processing information. Our introduction covered unsupervised as well as supervised machine learning, and we briefly mentioned many of the machine learning techniques and application areas.

We then presented a high level overview of the software tools available to the software developer, interested in using machine learning, and we decided to use the scikit-learn programming environment.

We then presented the series of steps we took using scikit-learn, to process a set of clinical data from the Memory Works clinic at Dublin City University. This dataset has only 120 subject visits/assessments, each with a groundtruth or eventual outcome for the subject, but we showed how the raw data needed to be cleaned and normalised before it could be used. We then applied a number of scikit-learn's in-build machine learning tools to build different machine classifiers and as part of this we were able to see which features, or what data from the clinical assessment, are most, and least useful. WE found that "age" is by far the most important feature, with "anxiety" being next important, grouped with a range of other related features.

In the next chapter we introduce another, much larger, dataset of clinical assessment, from the National Alzheimer's Coordinating Centre (NACC) in the US.

Chapter 4

The National Alzheimer's Coordinating Center (NACC) Dataset

4.1 Overview

As with all machine learning projects, the data used for training the machine learning algorithms is of paramount importance. It needs to be balanced, valid and true, and because the learning algorithms are automated, it needs to have volume. This work could be called a "big data" application where data is defined has having three "Vs", namely velocity, variety, and volume. These three aspects are defined as follows"

- **Velocity** refers to the changing nature of the data, how it changes over time and in our case this refers to how each case for each subject is different from the others, because people are different, therefore there is a constant change in our data;
- **Variety** refers to how data in a big data application is varied, it is unstructured rather than structured and it can be in the form of text, numbers, binary values etc. In our case the data captured from clinical sessions has such a varied nature;

- **Volume** refers to the huge volumes of data that can be processed in big data applications but in our case, because we are dealing with small numbers of individual people we fall short on this aspect of making our application a big data application.

We have already established in the previous Chapter that the data we have access to in the current Memory Works clinic database are not sufficient to train a machine learning classifier to assist with determining patient outcome. That dataset has only c.120 relevant instances, each corresponding to a single subject, when we examined it. Our task therefore is to source a dataset with the property that it has enough instances (people) and features (questions asked or diagnistic test results) that are pertinent to the target concept i.e dementia status assessment, and are suitable for training using a supervised machine learning algorithm. The sections below cover the provenance and contents of the dataset (NACC) that we have used.

4.2 The National Alzheimer's Coordinating Center (NACC)

The National Institute on Aging (NIA) of the National Institutes of Health (Bethesda, Md., USA) established the Alzheimer Disease Centers (ADCs) program beginning in 1984. Currently, 29 ADCs are funded by the NIA throughout the USA. The ADCs are similar in function to the Memory Works clinic established at Dublin City University, described in Chapter 1. They support a comprehensive approach to Alzheimer's disease (AD), including research on basic disease mechanisms, clinical and neuropathologic diagnosis and treatment, as well as educational initiatives for professional and lay audiences. Although the 29 ADCs share common components and features, each ADC developed its own set of unique research questions to ask of their subjects and methods for assessing and diagnosis. As a result, the content and administrative procedures for research protocols used to assess dementia at each ADC vary widely, as does the implementation of diagnostic criteria for mild cognitive impairment (MCI) and Alzheimer's Disease.

Using data collected from the 29 NIA-funded Alzheimer's Disease Centers (ADCs) across the United States, the National Alzheimer's Coordinating Center (NACC) has developed

and maintains a large relational database of standardized clinical and neuropathological research data. In partnership with the Alzheimer's Disease Genetics Consortium (ADGC) and the National Cell Repository for Alzheimer's Disease (NCRAD), NACC provides a valuable resource for both exploratory and explanatory Alzheimer's disease research. NACC data are freely available to all researchers under certain restrictions of dissemination and publication.

4.2.1 Description of the NACC Database

The NACC database is made up of three main research data sets defined as follows.

From 2005 and continuing to the present, ADCs have been contributing data to the Uniform Data Set (UDS), using a prospective, standardized, and longitudinal clinical evaluation of the subjects in the National Institute on Ageing's ADC Program.

In 2012, a new module was added to the UDS to collect detailed clinical information related to frontotemporal lobar degeneration (FTLD). The FTLD Module is voluntarily completed by ADCs.

Beginning in 1984 and ending with the 2005 implementation of the UDS, a brief, single-record descriptions of ADC subjects were collected retrospectively to form the Minimum Data Set (MDS).

The Neuropathology Data Set (NP) contains autopsy data for a subset of both MDS and UDS subjects. We should note that changes in diagnostic criteria and staining methods has limited the availability of some of this data for certain analyses.

The UDS is the dataset of interest to us and is available to researchers on request. The NACC also provide support for researchers in the form of direct contact via email and a document known as the Researcher's Data Dictionary[1]. The NP dataset might also be of interest for validation purposes with the machine learning algorithms we choose as it contains autopsy data that would confirm or reject any earlier clinician diagnosis. We did not, however use it for this purpose in the current work

[1]The Researchers' Data Dictionary is available at https://www.alz.washington.edu/WEB/rdd_uds.pdf

4.2.2 NACC Dataset Pre-processing

The first question we must address concerning the NACC dataset is which are the useful features from it we can use to construct a meaningful dataset to train machine learning algorithms. In this case, "meaningful" means that the features in NACC data map to features in the Memory Works clinic data, and we will return to this later.

The NACC dataset has approximately 105,400 instances with 530 features/variables associated with each instance. Because the dataset summarises data collected in a longitudinal study, many of the instances contain data for repeat visits by the same subjects. It is desirable that only the last visit by each subject is retained and used for the purposes of the machine learning exercise. When we removed the duplicate visit/assessment entries we were left with 32,594 unique instances.

In order to select the features to use we consulted an Alxheimer's Disease domain expert from the School of Nursing and Human sciences here in DCU. This domain expert suggested we use the risk factors recognised as being associated with developing Alzheimer's Disease and referred us to a paper [42] emanating from the InMINDD project, an EU FP7 funded project which aimed to identify the most accurate model of modifiable risk factors which can yield a personalised dementia risk profile. The goal reported in this referenced paper is to identify the major modifiable risk factors for dementia. The authors used two methods to identify the most important risk factors:

- literature review, followed by ...
- an online Delphi study which asked eight international experts to rank and weigh each risk factor for its importance for dementia prevention.

They found that there was was good agreement between modifiable risk factors identified in the literature review and risk factors named spontaneously by experts. After triangulation of both methods and re-weighting by experts, strongest support was found for the following risk factors:

- depression
- hypertension
- diabetes
- obesity

- physical inactivity
- alcohol
- hyperlipidemia
- smoking
- cardio-vascular disease

Each of these are modifiable Risk Factors, they are lifestyle factors that a subject can, with assistance, overcome and change so they are within a subject's control. There are other factors pertinent to the risk of developing Alzheimer's Disease, but are not modifiable, such as a persons age or sex. These were included in the dataset.

The task in pre-processing the NACC dataset was to map these identified risk factors to the features found in the NACC dataset. This involved extracting from the NACC dataset the features that recorded these risk factors. The mapping used is listed in Table 4.1.

Risk Factor	**NACC Data Dictionary Entry**
AGE ⇒	Calculated from (BIRTHYR-VISITYR)
SMOKING ⇒	TOBAC30/100, SMOKYRS, PACKSPER, QUITSMOKE
GENDER ⇒	derived feature IS_MALE from SEX
OBESITY ⇒	Calculated BMI $(WEIGHT(kgs)/HEIGHT^2(cms))$
DIABETES ⇒	DIABETES
MMSE ⇒	NACCMMSE
CARDIO-VASCULAR DISEASE ⇒	CVHATT
ALCOHOL ⇒	ALCOHOL
DEPRESSION ⇒	DEP
HYPERLIPIDEMIA ⇒	NOT AVAILABLE
PHYSICAL INACTIVITY ⇒	NOT AVAILABLE
OUTCOME ⇒	CDRGLOB

TABLE 4.1: Mapping from identified risk factors into the NACC data dictionary

The right hand side entries in Table 4.1 for the NACC data dictionary constitute the features we used in the machine learning classifier training set in our subset of NACC data. The last row of the Table (CDRGLOB) is the target variable, i.e. the label given to each instance by the examining clinician. This makes it the ground truth for the classifier. Some of the risk factors identified for us in [42] and listed above, were not directly present in the NACC dataset so there was no direct mapping, and these and are marked as not available. Accordingly they are not included as features in the training set

The NACC as an organisation, also records the subjects' scores for the Mini Mental State Examination (MMSE). The Mini Mental State Examination (MMSE) is a neuropsychological tool that can be used to systematically and thoroughly assess mental status. It is an 11-question measure that tests five areas of cognitive function: orientation, registration, attention and calculation, recall, and language. The maximum score is 30. A score of 23 or lower is indicative of cognitive impairment. The MMSE takes only 5-10 minutes to administer and is therefore practical to use repeatedly and routinely The MMSE is effective as a screening tool for cognitive impairment with older, community dwelling, hospitalized and institutionalized adults and is effective as a screening instrument to separate subjects with cognitive impairment from those without it.

The work by Joshi *et al.* [43] used MMSE scores in a machine learning experiment and reported good results. We have included the NACC variable for the total MMSE score in our experiments.

The iPython notebook we used to carry out the following experiments on the NACC data is similar to the one described in Chapter 3 on the Memory Works clinic data, and we will use the same presentation format of numbered cells as was used earlier in Section 1.3 for Memory Works clinic data processing. The experiments below detail the steps we took to extract the necessary instances and features from the original dataset received from the NACC.

```
In [1]:
import pandas as pd import numpy as np
    from sklearn.tree import DecisionTreeClassifier
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.linear_model import LogisticRegression
    from sklearn.neural_network import BernoulliRBM
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.svm import LinearSVC, SVC
    from sklearn.multiclass import OneVsRestClassifier
    from sklearn.naive_bayes import GaussianNB
    from sklearn import svm
    from sklearn.metrics import accuracy_score, f1_score, confusion_matrix
    from sklearn.cross_validation import cross_val_score, ShuffleSplit, StratifiedKFold
    from sklearn.preprocessing import Normalizer
    from sklearn.preprocessing import LabelBinarizer
    from sklearn.grid_search import GridSearchCV
import seaborn
```

```
In [2]: %matplotlib inline
In [3]: pd.options.mode.chained_assignment = None # default='warn'
```

```
In [3]: features = pd.read_csv('./nacc_original.csv', low_memory = False)

features = features.drop_duplicates(subset=['NACCID'], keep='last')

features = features[["VISITYR","VISITMO","BIRTHYR","BIRTHMO","DEP","SEX","TOBAC30",
                     "TOBAC100","SMOKYRS","PACKSPER","QUITSMOK","ALCOHOL","DIABETES",
                     "HYPERTEN","CVHATT","NACCMMSE ","HEIGHT","WEIGHT", "CDRGLOB"]]
```

The above cells [1] to [3] represent the usual module imports and other housekeeping. Cell [2] makes plots appear in the notebook as they are created and cell [3] deletes all patients' multiple visit data except the last one, in order to remove repeat visits to the ADC.

```
In [5]: features['DEP'].hist(bins=100) print(features.shape)

(32954, 19)

max size=0.90.901 - Feature Extractionfiles/01−FeatureExtraction91.png
```

in cell [5] we read in the original 105,400 instances. The repeat visit data is then removed and the features required for the experiments and described above are stored in the "features" dataframe where we initially have 22 columns for the features.

```
In [6]: features.head()

Out[6]:
        VISITYR VISITMO BIRTHYR    BIRTHMO  DEP     SEX TOBAC30 TOBAC100    SMOKYRS \
6       2013    2       1926        3       0       1   0       1           20
7       2014    5       1943        2       0       2   0       0           0
9       2007    6       1942        11      0       1   0       0           0
10      2008    7       1943        2       0       1   1       1           38
15      2013    5       1961        3       0       2   0       0           0

        PACKSPER    QUITSMOK    ALCOHOL DIABETES    HYPERTEN    CVHATT  NACCMMSE \
6       9           35          0       0           2           2       -4
7       0           888         0       0           1           0       97
9       0           888         0       0           0           0       30
10      1           888         0       0           0           0       30
15      0           888         0       0           1           0       30

        HEIGHT  WEIGHT  CDRGLOB
6       -4      -4      2.0
7       63      133     0.5
9       89      180     0.0
10      70      179     0.0
15      68      216     0.0
```

In Cell [6] we examine what features we have and a sample of (5) subjects and their features are presented for illustration.

```
In [7]: features['NACCMMSE'].hist(bins = 30)

Out[7]: <matplotlib.axes. subplots.AxesSubplot at 0x7fe9c7ae4d30>
max size=0.90.901 - Feature Extractionfiles/01 — FeatureExtraction131.png
```

Cell [7] looks at what the "mmse" scores look like, specifically those scores in the range 0 to +30, and the graph generated is shown in Figure 4.1, showing a gradual increase in distribution from 0 up to +30, with a few outlier results at the 100 mark..

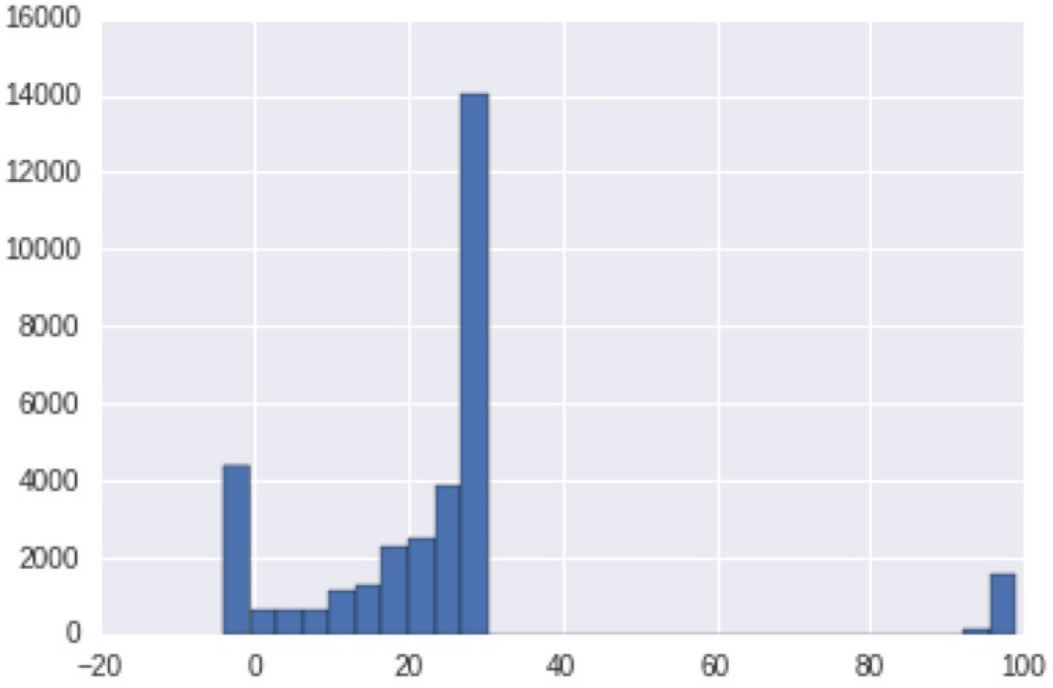

FIGURE 4.1: Plot of MMSE figures

```
In [8]:
    features = features[(features!=-4).all(axis=1)]
    features = features[features['HEIGHT'] > 0]

    features = features[features['WEIGHT'] > 0]
    features = features[features['QUITSMOK'] != 888]
    features = features[features['QUITSMOK'] != 999]

    features = features[features['NACCMMSE'] <= 30]
    features = features[~features['NACCMMSE'].isnull()]

    features = features[features['HEIGHT'] != 89]
    features = features[features['WEIGHT'] != 888]
```

In cell [8] we filter "weird" and unwanted values from the data, such as height and weight which are coded as negative values, and we eliminate other encoding for unknown values.

```
In [9]: features['HEIGHT'] = features['HEIGHT'] / 39.37
        features['WEIGHT'] = features['WEIGHT'] / 2.2046
```

Cell [9] converts height from inches to meters, and weight from lbs to kilos so we can-calculate BMI values.

```
In [10]: # Create features
    features["AGE"] = features["VISITYR"] - features["BIRTHYR"]
    features['is_male'] = (features['SEX']==1) * 1
    features['BMI'] = features['WEIGHT'] / (features['HEIGHT']**2)

    # check features
    features.head()

Out[10]:
    VISITYR VISITMO BIRTHYR BIRTHMO DEP SEX TOBAC30 TOBAC100    SMOKYRS \
26  2012    10      1950    6       1   2   0       1           1
29  2012    9       1928    10      0   1   0       1           30
39  2013    1       1045    8       0   2   0       1           2
45  2013    1       1936    9       0   1   0       1           26
59  2009    12      1944    4       0   2   0       1           15

    PACKSPER ...    DIABETES    HYPERTEN    CVHATT  NACCMMSE    HEIGHT \
26  1               0           0           0       8           1.600
29  1               1           1           2       16          1.828
39  2               0           0           0       23          1.625
45  5               0           1           0       29          1.778
59  2               0           0           0       28          1.574

        WEIGHT      CDRGLOB     AGE is_male     BMI
26      87.090629   1.0         62  0           34.01
29      74.389912   1.0         84  1           22.24
39      63.049986   1.0         68  0           23.85
45      65.771569   0.5         77  1           20.80
59      48.534882   0.0         65  0           19.57

[5 rows x 22 columns]
```

Cell [12] creates the features to be used in the training and learning.

```
In [13]: features['BMI'].hist(bins=100)
         print(features.shape)

(9634, 22)
max size=0.90.901 - Feature Extractionfiles/01 – FeatureExtraction201.png
```

In cell [13] we check to ensure that BMI values look reasonable and these are plotted in Figure 4.2, showing a good distribution of values, peaking in the mid-20s.

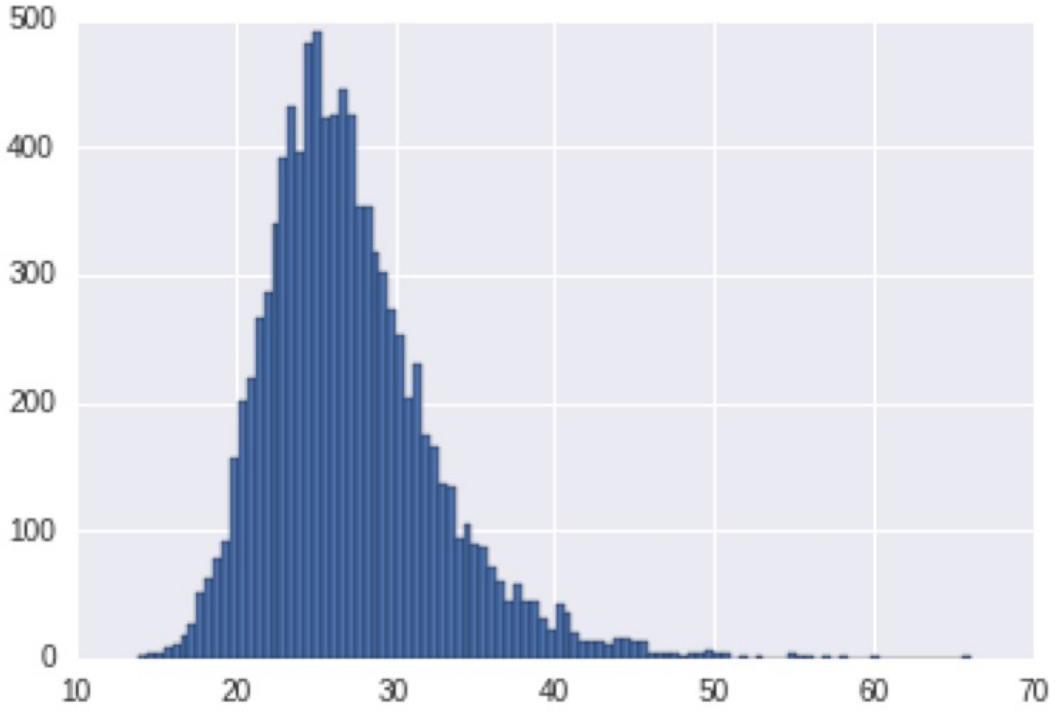

FIGURE 4.2: Plot of BMI values

```
In [17]: def score2string(n):
             if n == 0: return 'no'
             elif n == 0.5: return 'questionable' elif n == 1: return 'mild'
             elif n == 2: return 'moderate'
             elif n == 3: return 'severe'
             else: return 'not found'

    features['CDRGLOB'] = features['CDRGLOB'].apply(score2string)
```

In cell [17] we rename some of the categories, replacing the encodings with the class of dementia, if present

```
In [18]: features.columns print(features.shape)17 creates

(9634, 22)
```

In cell [18] we once again check that we have the required features.

```
In [20]:    features = features.drop(['VISITYR', 'VISITMO', 'BIRTHYR', 'BIRTHMO',
                                     'HEIGHT','WEIGHT', 'SEX']
            print(features.shape)

(9634, 15)
```

In cell [20] we remove some unwanted features, reducing from 22 to 15.

```
In [17]: features.to_csv('./nacc_extracted.csv', index=False)
```

Finally, pre-processing the data is complete and we create the CSV file with the processed and extracted data. This is the set of instance that will be used in the supervised machine learning experiments and contains 9,634 instances or subjects, each with 15 essential features.

4.3 Summary

In this Chapter we introduced a clinical assessment dataset provided to us by the US National Alzheimer's Coordinating Centre (NACC). We described the data, how and from where it was collected, and how a set of dementia risk factors, derived from work in the IN-MINDD project, could be successfully mapped to fields in the NACC data dictionary.

We then presented a series of iPython commands necessary to re-format and to structure the NACC data so it could operate within the scikit-learn programming environment. THis involved data clearning and alignment, and preparing inputs so the clinical data parameters can be used as features in the machine learning process.

In the next chapter we will present our initial set of experiments on that NACC data.

Chapter 5

Machine Learning on the NACC Data

5.1 Introduction

The clinical data from the NACC repository at the University of Washington in the US and described in the previous Chapter, was received in the form of a comma separated values (CSV) file containing 105,400 instances, each with 530 features. This clinical data was collected by the US National Institute for Aging funded Alzheimer's Disease Centers (ADCs). These ADCs are similar to the Memory Works clinics described in Chapter 1. There are 29 of these ADCs throught the US. Subjects concerned about dementia can visit one of the ADCs where they are assessed for potentially different stages of Alzheimer's disease. Subjects may visit the ADCs on multiple occasions and data is recorded on each occasion. The data collected from these visits are standardised and sent to the University of Washington to be stored in a repository. This anonymised data is available for research purposes.

We submitted a request for the most recent set of data and the request was granted. For the purposes of our experiment we needed to extract a reasonably-sized dataset containing only the pertinent features which map to the Memory Works clinic data, to use for our machine learning work. The definition of the *pertinent features* was described earlier in Chapter 3. The first pre-processing task was to eliminate the longitudinal aspect of the data by deleting instances recorded for repeat visits by a subject, retaining

only the record of the most recent visit. We then needed to derive some features not directly contained in the data, such as the subjects' age at the time of the visit, and their body mass index (BMI) calculated from their weight in kilos and their metric height. The Boolean value "is_male" was also derived from the "sex" variable.

The pre-processing we carried our is described in detail within the IPython notebook included in Chapter 4. The NACC variable CRDGLOB records the Clinical Dementia rating resulting from the subjects' visit. In the dataset this is encoded as a continuous nummeric value with the allowable codes below:

- 0.0 = No impairment
- 0.5 = Questionable impairment
- 1.0 = Mild impairment
- 2.0 = Moderate impairment
- 3.0 = Severe impairment

This is the target, or class variable, in the supervised machine learning exercise that we are aiming to estimate. For the purposes of the experiment, it was necessary to transform these values into a categorical form, with text replacing the continuous encoding representation.

This previous Chapter, Chapter 4, described how we used the Python Pandas library to extract the data required for our machine learning experiment and to perform pre-processing. Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. The extracted data was written to a csv file. This chapter describes how we used the data extracted from the NACC data to train various machine learning algorithms. In a way, in this chapter we report intermediate results and set up the configuration where, in the next chapter, we present and discuss the main results of the thesis.

5.2 Using Scilit-learn to Generate Supervised Machine Leaning Classifiers

As with previous chapters, we present the iPython notebook containing the code we used which is detailed below and presented as a series of numbered cells. The first cell [1] imports the required modules

```
In [1]: import pandas as pd
        import numpy as np

In [3]: from sklearn.tree import DecisionTreeClassifier
    from sklearn.neighborsDecisionTreeClassifier import KNeighborsClassifier
    from sklearn.svm import LinearSVC, SVC
    from sklearn.linear_model import LogisticRegression
    from sklearn.naive_bayes import GaussianNB
    from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier
    from sklearn.ensemble import VotingClassifier
    from sklearn.dummy import DummyClassifier
    from sklearn.multiclass import OneVsRestClassifier
    from sklearn.cross_validation import cross_val_score, StratifiedKFold, KFold
    from sklearn.preprocessing import LabelBinarizer
    from sklearn.pipeline import Pipeline

import pandas as pd

import seaborn as sns
```

These cells [1] and [3] contain boilerplate code for importing the modules from scikitlearn required for classification and cross validation. Note the powerful nature with which we can use scikitlearn to invole really complex machine learning classifiers like k-nearest neightbours, decision trees, support vector machines, random forest and extra trees classifiers, etc., and with which we can invoke libraries to perform n-fold cross validation.

```
In [4]:
    features = pd.read_csv('./nacc_extracted.csv')
```

In cell [4] we read in the the file generated by the IPython notebook used for extracting the data from the NACC dataset and performing pre-processing and data cleaning, as described in Chapter 4.

```
In [5]: features.head()

Out[5]: TOBAC30 TOBAC100 SMOKYRS PACKSPER QUITSMOK ALCOHOL DIABETES \
0       0       1        1       1        18       0       0
1       0       1        30      1        50       0       1
2       0       1        2       2        22       0       0
3       0       1        26      5        38       0       0
4       0       1        15      2        31       0       0

        HYPERTEN    CVHATT  NACCMMSE    HEIGHT  WEIGHT  CDRGLOB     AGE \
0       0           0       8           1.60    87      mild        62
1       1           2       16          1.82    74      mild        84
2       0           0       23          1.62    63      mild        68
3       1           0       29          1.77    65      questionable 77
4       0           0       28          1.57    48      no          65

        is_male BMI
0       0       34.011137
1       1       22.242310
2       0       23.859200
3       1       20.805250
4       0       19.570478
```

In cell [5] above, we print the first five rows, just for a visual check

```
In [6]: y2 = features['CDRGLOB']
            y = pd.get_dummies(y2)
            X = features.drop(['CDRGLOB'], axis=1)
```

This cell, [6], creates the training set matrix, X, and the category vector, y used by the classifiers.

```
In [7]:
        skfold = StratifiedKFold(y=y2, shuffle=True, n_folds=10)
```

In cell [7] we create the stratified K-Folds cross validation iterator. This provides train/test indices to split data into train/test sets, on our case above we are using 10-fold valudation. The cross-validation object is a variation of KFold that returns stratified folds. The folds are made by preserving the percentage of samples for each class.

```
In [8]:
    dc_skfold_score = cross_val_score(DummyClassifier(strategy='stratified'),
                                      X, y2, scoring='accuracy', cv=skfold)
```

in cell [8] we create a dummy classifier used for comparison purposes with other classifiers. The other classifiers must perform better than this baseline. In the next series of cells, [9] to [14] we create the real classifiers and perform the cross-validation scoring. The classifiers we are using are decision trees (cell [9]), random forest ensemble [10], extra trees [11], k-nearest neighbours [12], support vector machine [13] and GLM regression [14].

```
In [9]: ovr_dt = OneVsRestClassifier(DecisionTreeClassifier())
dt_kfold_score = cross_val_score(ovr_dt, X, y, cv=10, scoring='accuracy')
dt_skfold_score = cross_val_score(DecisionTreeClassifier(), X,y2 scoring='accuracy',
cv=skfold)

In [10]: ovr_rf = OneVsRestClassifier(RandomForestClassifier())
rf_kfold_score = cross_val_score(ovr_rf, X, y, cv=10, scoring='accuracy')
rf_skfold_score = cross_val_score(RandomForestClassifier(), X, y2, cv=skfold)

In [11]: et_skfold_score = cross_val_score(ExtraTreesClassifier(), X, y2, cv=skfold)

In [12]: ovr_knn = OneVsRestClassifier(KNeighborsClassifier())
knn_kfold_score = cross_val_score(ovr_knn, X, y, cv=10, scoring='accuracy')
knn_skfold_score = cross_val_score(KNeighborsClassifier(), X, y2, cv=skfold)

In [13]:
ovr_svm = OneVsRestClassifier(LinearSVC(class_weight='balanced'), n_jobs = -1)
svm_kfold_score = cross_val_score(ovr_svm, X, y, cv=10, scoring='accuracy')
svm_skfold_score = cross_val_score(LinearSVC(class_weight='balanced'), X, y2,
cv=skfold, scoring='accuracy')

In [14]: lr_skfold_score = cross_val_score(LogisticRegression(), X, y2, cv=skfold,
scoring='accuracy')

nb skfold score = cross val score(GaussianNB(), X, y2, cv=skfold, scoring='f1 weighted')
```

The wall time to complete the SVM processing including training and 10-fold cross validation, for example, on a standard DELL desktop computer running Windows was 2 minutes and 38 seconds.

```
In [16]: # All results
        scores = [dt_kfold_score, dt_skfold_score,
                    rf_kfold_score, rf_skfold_score, knn_kfold_score,
        labels = ['tree 1', 'tree 2', 'forest 1','forest 2','knn 1',
                    'knn 2', 'svm 1','svm 2']
        scores = pd.DataFrame(scores, index=labels).T
        sns.boxplot(data=scores)

Out[16]:
        <matplotlib.axes. subplots.AxesSubplot at 0x7681090>
```

Having set up the environment and calculated the performances for different classifiers, we now first plot all the results, with standard cross-fold validation results and then with the stratified-cross fold validation. The graph in Figure 5.1 shows the results for the following classifiers . . . decision tree, random forest, k-nearest neighbour and support vector machine. For each classifier we present two results (tree1 and tree2, forest1 and forest2, etc.). Each entry is presented as a boxplot, where the line in the middle of the box indicates the median value of the ROC AOC, and the bottom of the rectangle drawn around that median represents the first and the top of the rectangle represents the third quartiles, respectively. The "whiskers" around the rectangle represent the maximum and minimum data values. This will help us to interpret the boxplot presentation, and the "tighter" the boxplot, the more reliable the estimation, i.e. the less the variation across the n (10) folds of evaluation. The first entry in the graph in each case is for the score of the classifier as measured using *standard* k-fold cross validation, the second is the score using *stratified* cross-validation.

From Figure 5.1 we can see that the random forest classifier with stratified cross-validation is clearly the best performer in terms of ROC-AUC score. The score is over 0.5 and there is little variation in the results, as shown in the boxplot, in particular the horizontal area around the result which indicates the amount of variance. The k-nearest neighbour classifier, also with a stratified cross-validation is second in terms of performance, with a median value of just over 0.5. For all classifiers, performance using the stratified cross-validation is better than standard.

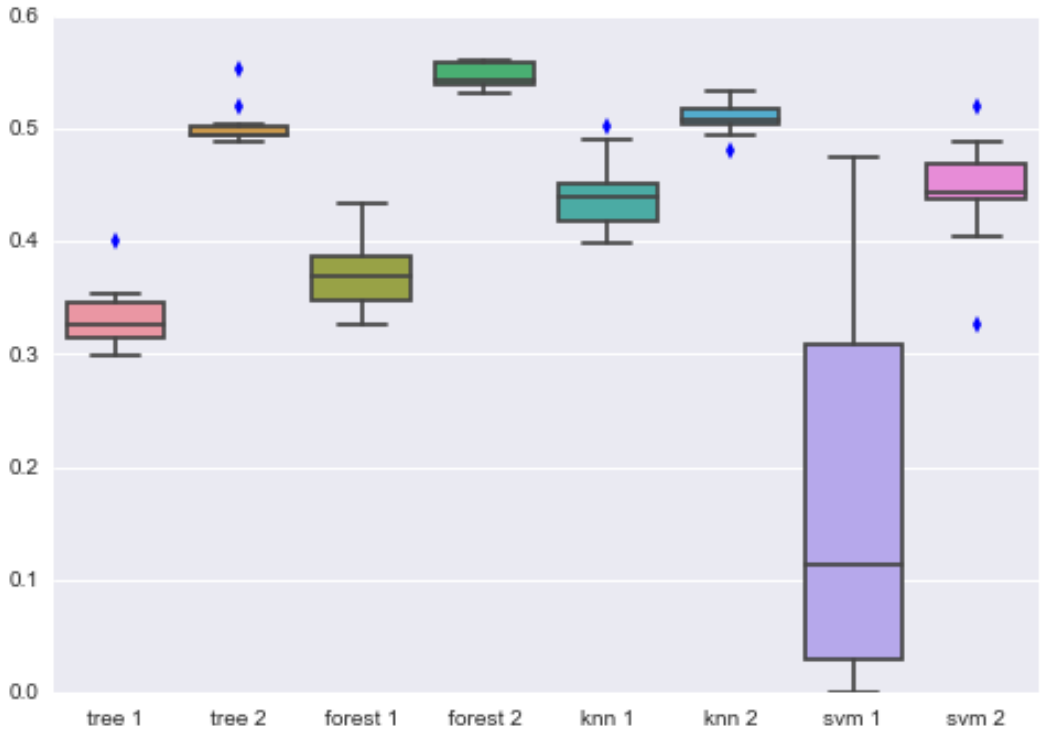

FIGURE 5.1: Boxplot with all ROC-AUC scores

```
In [17]: # Just stratified
        scores = [dc_skfold_score, dt_skfold_score,
                  rf_skfold_score, knn_skfold_score, svm_skfold_score,
                  lr_skfold_score, nb_skfold_score, et_skfold_score]
        labels = ['dummy', 'tree', 'forest', 'knn', 'svm', 'logistic',
                  'bayes', 'extra trees']

        scores = pd.DataFrame(scores, index=labels).T

        sns.boxplot(data=scores)
```

In cell [17] we generate an analysis and plot with only the stratified results This plot is for the classifiers : dummy, decision tree, random forest, k-nearest neighbour, support vector machine, logistic regression, naive Bayes and extra tree. The result is measured using stratified k-fold validation where the value of k is set to 10. The results are presented in Figure 5.2.

Figure 5.2 shows that the best performing classifier is the logistic regression classifier, with close second performance in terms of ROC-AUC values from random forest, naive Bayes and extra tree. Based on this result, this is the classifier we will use in the remainder of the work, where to attempt to establish an estimate of the minimum

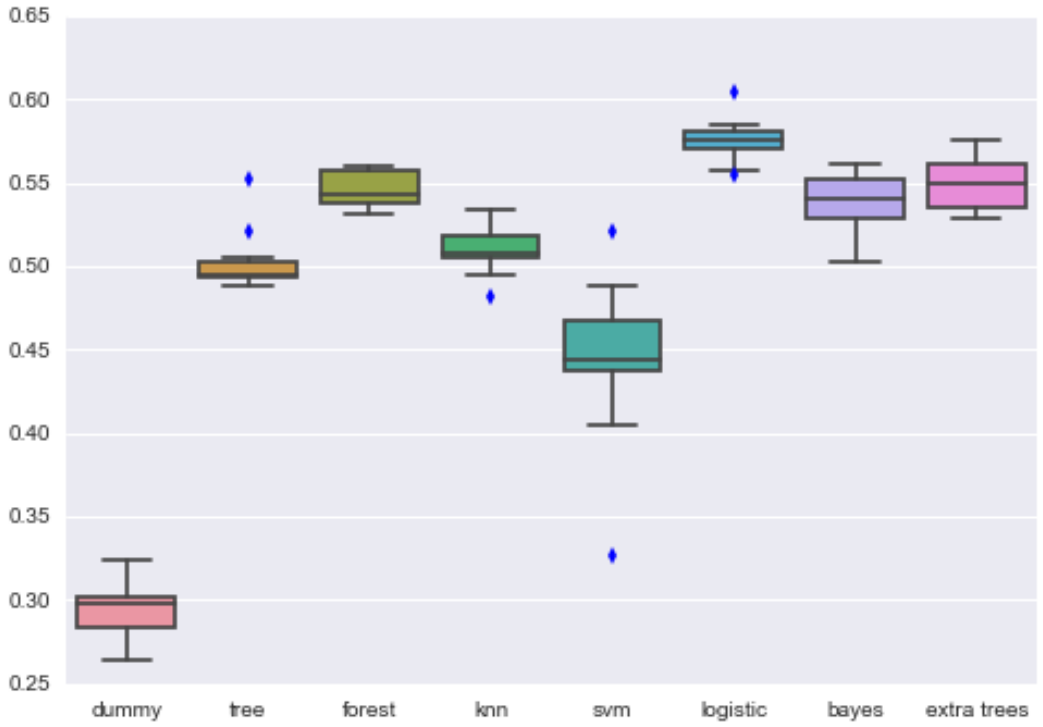

FIGURE 5.2: Boxplot with AUROC scores from stratified k-fold cross-validation

number of instances we need to train a classifier for screening subjects for possible Alzheimer's Disease.

5.3 Machine Learning Techniques for Diagnosis of Alzheimer's Disease

In the previous section we have shown that we are able to train several classifiers using the full dataset of over 9,000 instances and tested their accuracy using both standard k-fold and stratified k-fold cross validation. We found that the best performing classifier as measured by stratified k-fold cross validation was the logistic regression algorithm. Using this result we can say that it is possible to train a supervised machine learning algorithm to assist in the screening/diagnosis of subjects for Alzheimer's Disease. We will use this classifier when addressing the next question concerning the optimal number of instances we need for the task but first, let us see have we answered the first of our research questions, which can be re-stated here as ...

1. **RQ1:** What is the accuracy with which a subject's diagnosis of dementia can be predicted based on a set of clinical data for that subject, compared to patterns mined from a dataset of previous diagnoses ?

Clearly, given the results presented in Figure 5.1 and Figure 5.2, we have answered this question, with an ROC AUC value of about 0.58 for the best-performing classifier. Later in this chapter we shall return to the question of what exactly these ROC AUC values mean in practice, but for now we can move on to the second of our research questions.

5.4 The Number of Instances of Clinical Data Required

As described in section 3.3 above, we were not able to use the c. 120 instances in the Elevator database to train a classifier so that it could be successful in prediction tasks for Alzheimer's Disease diagnosis. It appeared that there were an insufficient number of instances to do this. This raised the question of how many instances are necessary to achieve this ? This was formalised earlier as ...

2. **RQ2:** Clinical assessment data for what number of subjects, is required in order to automatically mine patterns and make a reliable diagnosis for a new, as yet unseen, subject ?

The previous section above here answered our first research question, in that it appears feasible to carry out this task using machine learning, and we will now use the results of our experiments to attempt to answer the second research question. We have already found that the Logistic Regression classifier was the best performer for prediction, so we will use this in experiments to test the hypothesis that there is an optimal number of instances that we can use to train the algorithm.

The approach we used was to plot a graph of the scores resulting from using stratified k-fold cross validation using the Logistic regression classifier on training sets of varying and increasing sample sizes taken from the full dataset. We start with a sample size of 120 instances, as this was the number in the Elevator database. For the cross validation, k is set to 20. For each subsequent iteration, we generate a sample size with 100 more instances up to a sample of 9,000 from the original dataset extracted from the NACC set.

The resulting score is held in the score python list. The final score object is then placed in a pandas Dataframe (scores) for graphing. The code from the IPython notebook below implements this approach. The last line below generates the graph. The graph is shown in Figure 5.3 below.

```
10 Number of instances required

In [129]: %%time
        sizes = range(120, 9600, 100)
        scores = []

        for size in sizes:
            idx = X.sample(size).index
            X_sample = X.loc[idx]
            y_sample = y2.loc[idx]
            kfold = KFold(size, n_folds=20)
            skfold = StratifiedKFold(y_sample, n_folds=20)
            score = cross_val_score(LogisticRegression(), X_sample,
            y_sample, scoring='accuracy',
            cv=skfold, score = [(size, j, i) for i, j in enumerate(score)]
            scores += score
        scores = pd.DataFrame(scores, columns=['sample size', 'score', 'fold'])

In [130]:
        sns.tsplot(data=scores, time='sample size', value='score', unit='fold')
```

The x (horizontal) axis on the graph is the number of instances, while the vertical y axis is the prediction score in percent correct as measured by the stratified k-fold cross validation. We are looking for the point where the scores are between 55 - 60 % and where the graph starts to show signs of stability i.e where there is less variability.These are the values shown in the boxplot above for this particular classifier. We can see that, at the lower end of the sample size range, the score is both below 50%, and unstable. It is only when the number of instances passes about 3,500 that it starts to take on the characteristics we are looking for, so we estimate that this is roughly the minimum number of instances we would need to get a reasonably accurate predictive Supervised machine learning model for our purposes.

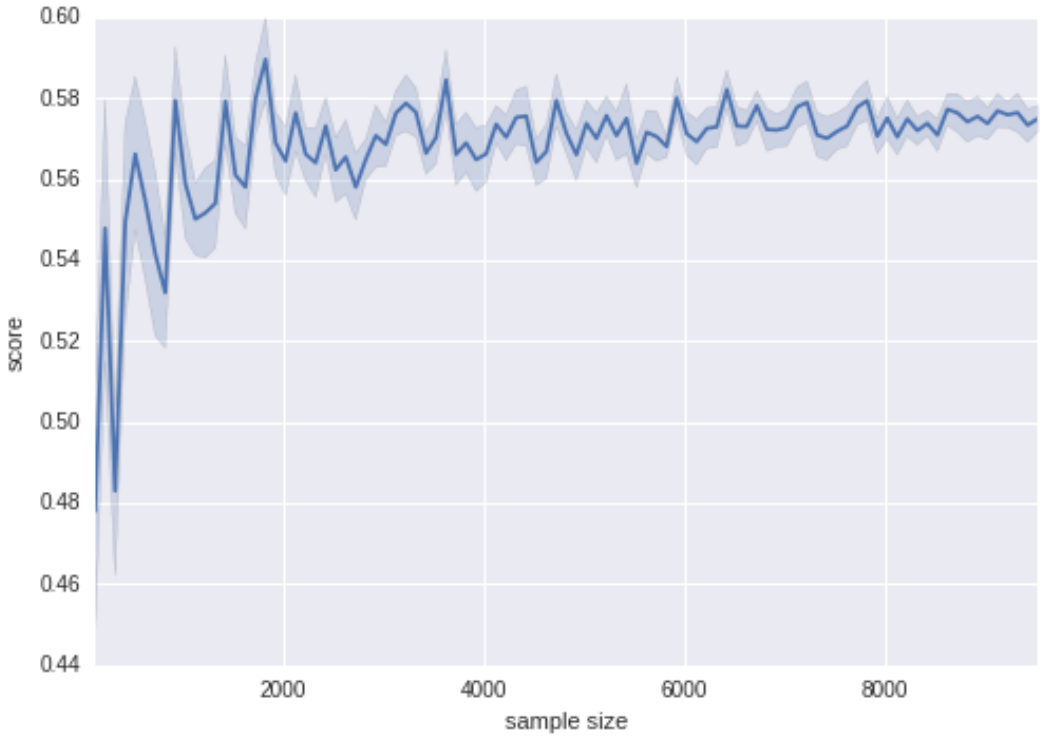

FIGURE 5.3: Graph of score for training sets on increasing size

```
In [26]:
        small_sample_scores =scores[scores['sample size'] < 1000]

In [27]:
        sns.tsplot(data=small_sample_scores, time='sample size',
        value='score', unit='fold')

Outv[27]:
        <matplotlib.axes. subplots.AxesSubplot at 0x7fbdd014d588>
```

In the graph below in Figure 5.4, created by cells [26] and [27] above, we zoom in on score results for samples at the lower end of the range i.e from our starting number of samples at 120, up to 1,000 samples. From that we can see that the scoring begins at just under 0.50, but gradually rises to over that, albeit with a large amount of variation. It is obvious that as we increase the number of samples, the score rises. Thus we confirm that we can increase the scoring accuracy for prediction by adding more samples up to a minimum level, and this is in fact the fourth of our research questions, re-stated below for clarity:

4. **RQ4:** What does the tradeoff between accuracy of diagnosis prediction vs. number of subjects in the training dataset, look like and is there a "sweet spot" in that tradeoff ?

We can now regard this research question as being answered.

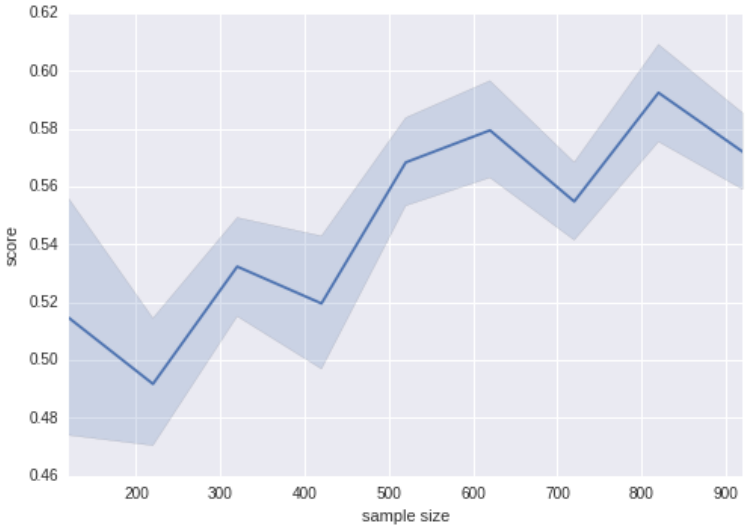

FIGURE 5.4: Graph of score for training sets at lower end of range

5.5 Model Evaluation

The goal of model evaluation, which is what we are doing throughout this thesis, is to help to estimate how well a particular model of the data will generalise to new data so that we can choose between different models. For this, not only do we need an evaluation procedure such as train/test split or cross-validation, but we also need an evaluation metric in order to quantify performance of different models.

A very simple metric is classification accuracy, which measures how often a model correctly predicts the class of an instance in the validation set. However, classification accuracy does not take into account the underlying distribution of the test set data values and sometimes can mask poor performance where there is a skew or imbalance in the distribution of values, so the model appears to be highly accurate when is not actually.

Classification accuracy alone obscures two critical pieces of information: the underlying distribution of the response values, and the "types" of errors that the classifier is making.

As such, classification accuracy alone does not give a clear picture of how a classifier is actually performing.

Two other metrics metrics in use for model evaluation in the literature for machine learning are the confusion matrix and the AUROC. Performance metrics such as Precision, Recall and F1 can be calculated from the confusion matrix but we do not use these as they are not applicable to this work since they depend on a ranking whereas we assign classification outputs independent of each other.

AUROC is the area under the ROC (receiver operating characteristic) curve. Its name comes from the fact that is was first used by British radar engineers during World War 2 to tune radar equipment to distinguish between incoming German planes and other airborne objects such as flocks of birds.

The Receiving Operating Characteristic, or ROC, is a visual way for inspecting the performance of a classification algorithm. In particular, it is used to compare the rate at which a classifier is making correct predictions (True Positives or TPs) and the rate at which a classifier is making false predictions (False Positives or FPs). When talking about True Positive Rate (TPR) or False Positive Rate (FPR) we are referring to the definitions below:

$$TPR = TruePositives/(TruePositives + FalseNegatives)$$

$$FPR = FalsePositives/(FalsePositives + TrueNegatives)$$

True Positives and True Negatives are also referred to as Sensitivity and Specificity. What we are measuring here is the trade-off between the rate at which one can correctly predict something, with the rate at which the classifier predicts something that doesn't happen.

A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Each prediction result or instance of a confusion matrix represents one point in the ROC space.

The best possible prediction method for a machine learning classifier would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). This (0,1) point

is also called a perfect classification. A completely random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners (regardless of any skew in the positive and negative base rates). An intuitive example of random guessing is a decision made by flipping coins (heads or tails). As the size of the sample increases, a random classifier's ROC point migrates towards (0.5,0.5). We will attempt to illustrate how the ROC space is plotted by presenting some examples. Figure 5.5 below is one example. It is for a classifier with an equal number of TPRs and FPTs and shows that the classifier is no better than guessing when predicting an outcome.

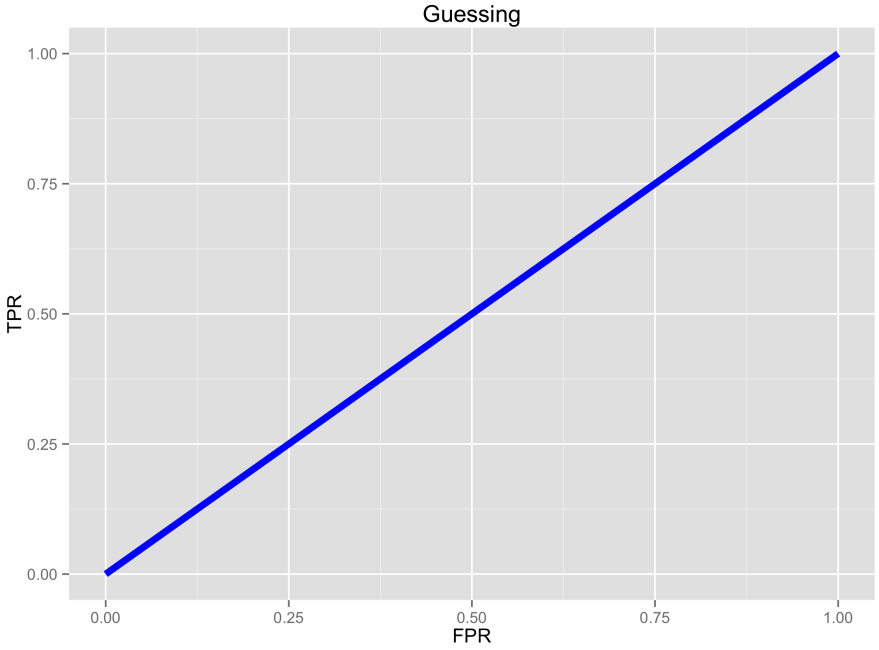

FIGURE 5.5: Graph of AUROC where model is no better than guessing

The next graph in Figure 5.6 shows AUROC scores for a bad performing classifier. All or most of the scores are below the diagonal.

In the next graph in Figure 5.7, the scores are all above the diagonal, which indicates the classifier is performing moderately well, if not spectacular. This is that approximate score for our logistic regression classifier,i.e in the 55 - 60% range

This graph shown in Figure 5.8 shows the scores for a classifier which is performing pretty good. All of the scores are above the disagonal, the gradient in the curve is rising and it tends towards the upper left corner or coordinate (0,1) had of the ROC space.

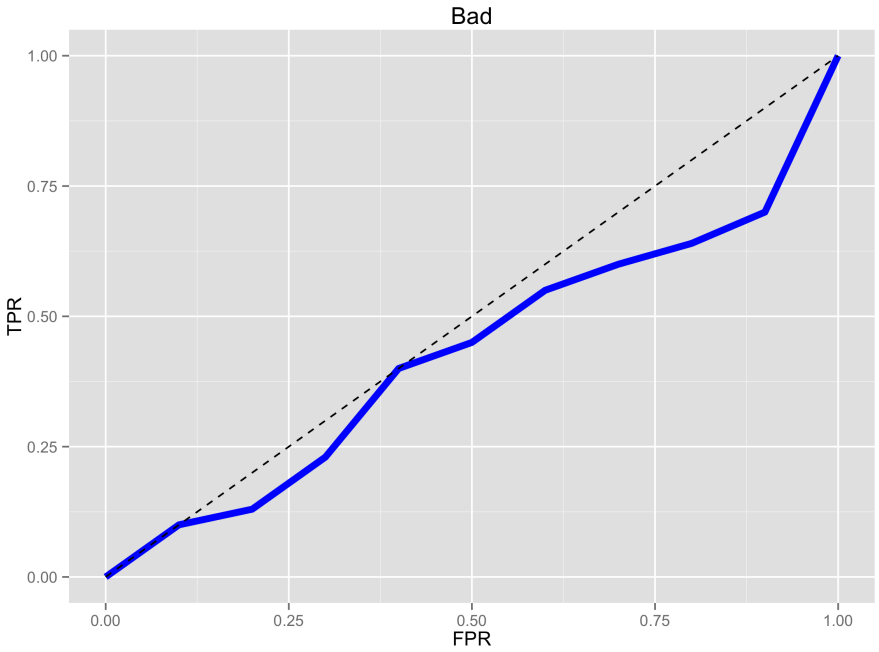

FIGURE 5.6: Graph of AUROC where model is not performing well at all

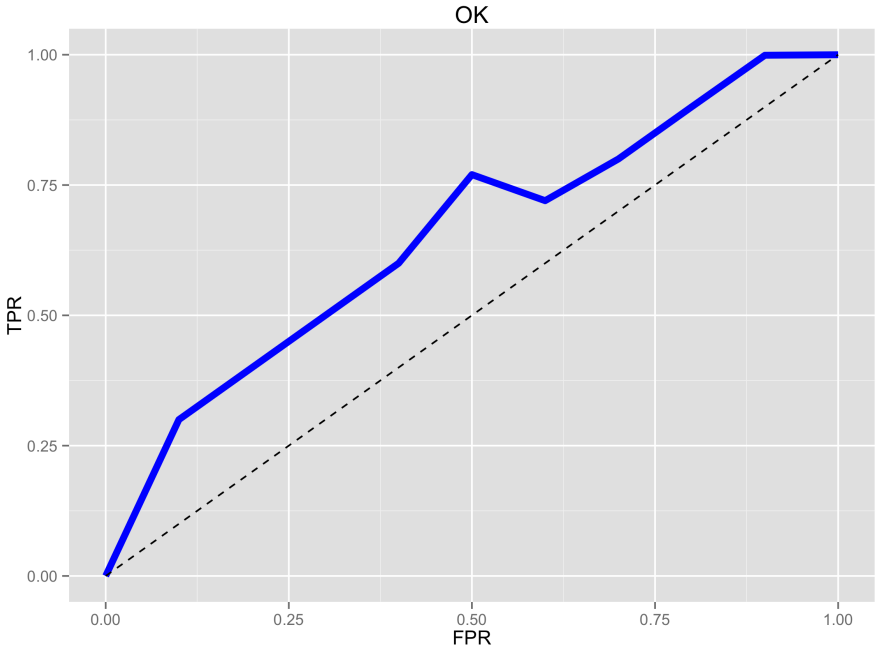

FIGURE 5.7: Graph of AUROC where model is OK, just about

Finally, This graph in Figure 5.9 represents the area under the curve for the very well performing model. The area is measured as the proportion of the area under the ROC curve as a fraction of te overall area in the graph, and the closer the proportion of the area is to a total value of 1.0, the better the model is performing.

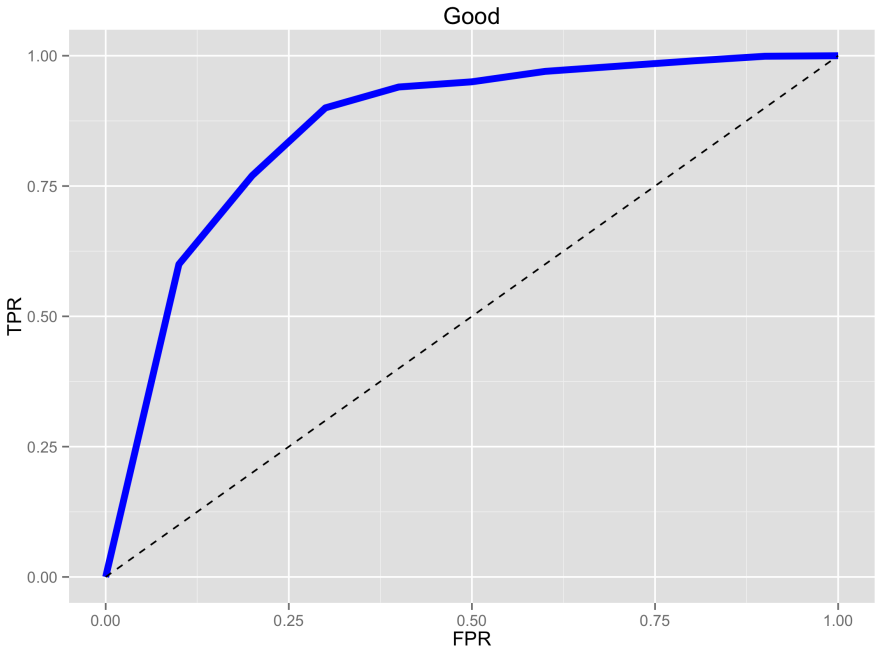

FIGURE 5.8: Graph of AUROC where model is performing very well

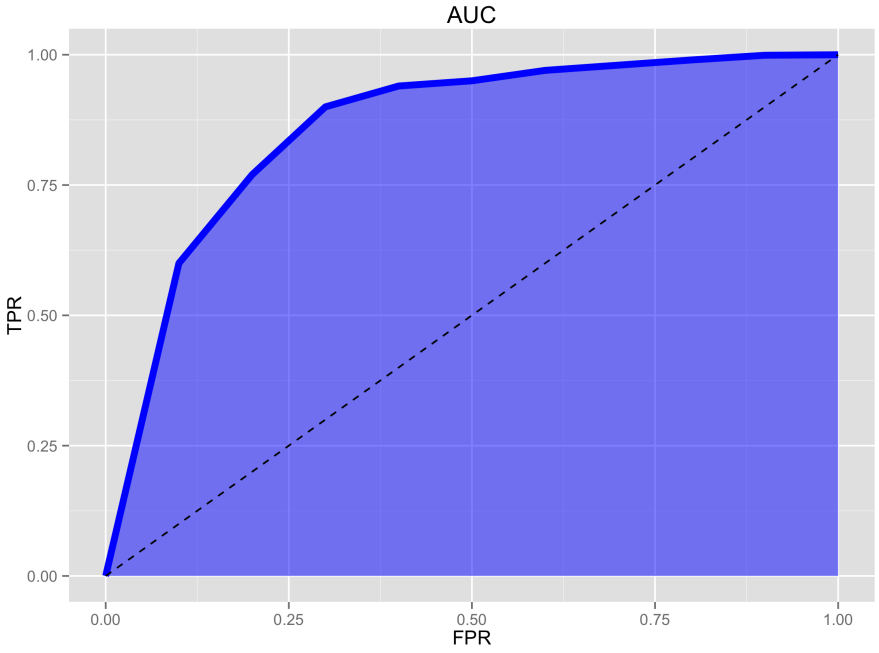

FIGURE 5.9: Graph of Area under curve for very good model

Throughout this thesis we have used AUROC as a measure of performance of a classifier, when experimenting on data from both the Memory Works clinic and the NACC data. Clearly, since we are not getting AUROC values close to 1.0 we are not getting perfect classifiers, and it would be astounding if we did, but the top values we are achieving, in the 0.55 to 0.60 range, point to classifiers which are very useful.

5.6 Importance of Features

The last remaining unanswered research question we have in the thesis is the following ...

1. **RQ3:** In carrying out a prediction of diagnosis outcome for a subject, which clinical features are more, or less, important ?

Earlier in the thesis we did some work on the data from the Memory Works clinic and found that the most important and discriminating feature was age and we showed this in Figure 3.7 from Chapter 3. But what can you say about this question and the NACC data feature importance and does the same hold true for this larger dataset ? We will now repeat that exercise carried out in Figure 3.7, for the features we used in the NACC dataset.

As we reported earlier in Chapter 3 the ExtraTreesClassifier classifier can handle multiple target variables, and can also rank the importance of each feature. So, in cell [28] from our iPython notebook we generate a Pandas series object with the features ranked in descending order.

```
In [28]:
    forest = ExtraTreesClassifier(n_estimators=250,
                                  random_state=0)
    forest.fit(X, y2)
    importances = forest.feature_importances_importances =
        pd.Series(importances, index=X.columns)
    importances.sort(ascending=False)
    importances

Out[28]:
    NACCMMSE    0.299305
    AGE         0.087329
    BMI         0.086628
    WEIGHT      0.084773
    QUITSMOK    0.082846
    SMOKYRS     0.080604
    HEIGHT      0.076994
    PACKSPER    0.057916
    HYPERTEN    0.040369
    DIABETES    0.026174
    CVHATT      0.017759
    DEP         0.017493
    ALCOHOL     0.016839
    is_male     0.016249
    TOBAC30     0.005142
    TOBAC100    0.003580

    dtype: float64

In [24]:
        importances.plot(kind = "bar")
```

We print the values and the graph in Figure 5.10 shows a bar chart of the features in terms of their importance. From this we can see that MMSE is the most significant feature, which we would expect, is followed by age and BMI. This provides our answer to research question 3 (RQ3). The less important features at the rightmost end of the bar chart could be dropped in any further iterations of our experiments.

5.7 Summary

In this Chapter we discussed how we applied scikit-learn machine learning algorithms to the selected features from the NACC dataset to generate several classification models. We found that the best performing model was Logistic Regression, which despite it's name is a classification model. When we ran stratified cross validation on this model, parameterised to give us ROC-AUC scoring, we found that it performed reasonably, but

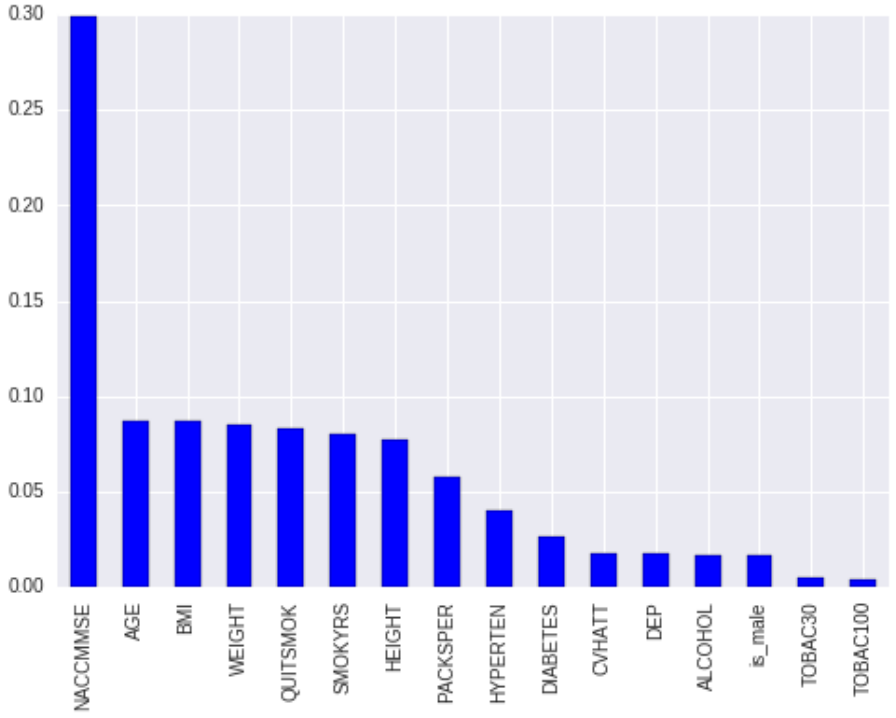

FIGURE 5.10: Relative Importance of Training Features in the NACC Dataset

not spectacularly well. When we graphed it, the graph would be similar to Figure 5.7 above.

We also described in the Chapter what an AUROC is used for, and included various graphs to illustrate the results of using AUROC scores for models at different performance levels. Finally we identified and extracted the most important features used in training the models, and presented these.

Chapter 6

Conclusions & Discussion on the results of the experiments to answer the Research Questions

Before we present some conclusions about the results from the experiments in the thesis we need to start with the assumptions we have made about the NACC dataset and the features selected from the dataset on which to base our machine learning implementation. We must ask ourselves if the results we obtained are reliable. To answer that question we state our assumptions about the data used and the features we used from the NACC data to train the scikit-learn algorithms.

To quote from the NACC website :

"The National Alzheimer's Coordinating Center was established ... to facilitate collaborative research. Using data collected from the 29 NIA-funded Alzheimer's Disease Centers (ADCs) across the United States, NACC has developed and maintains a large relational database of standardized clinical and neuropathological research data. In partnership with the Alzheimer's Disease Genetics Consortium (ADGC) and the National Cell Repository for Alzheimer's Disease (NCRAD), NACC provides a valuable resource for both exploratory and explanatory Alzheimer's disease research. NACC data are freely available to all researchers."

That defines the setup and function of the NACC so it appears at face value that it is a viable dataset for us to use.

The full database comprises several standardized clinical and neuropathology data sets, all of which are freely available to the research community. The University of Washington's Data Science Department is well respected, so we would expect that the data accurately reflects the data collected from the 29 ADC's and are relevant to the subject at hand, i.e research in all aspects of Alzheimer's disease. Assuming this is the case, we now state our assumptions about the features we extracted from the NACC data to use for training purposes.

We based our decision on what features were optimal to use for training purposes in our experiments, on our experiences with the Elevator project and the Memory Works clinic dataset described earlier in Section 1.3. We assumed that the features to use in machine learning how to predict diagnosis were the modifiable risk factors associated with contacting Alzheimer's Disease. Besides these, we added two other features that have a strong correlation with the onset of Alzheimer's Disease, namely age and gender. We believe it is a reasonable assumption that these were the correct features to use. Further research might prove us wrong, and that other features within the NACC dataset would give more accurate and/or higher scores but because our motivation in using the NACC dataset is to map our findings back to the Elevator project and the Memory Works clinic data, even if there are more useful NACC dataset features, unless they are in the Memory Works clinic dataset, they are of no value to us for the Elevator project. Our analysis of the most important features described in Chapter 5 indicates that some of the features could be dropped from the training set

In our experiments we have established that supervised machine learning algorithms can be used to predict whether a subject is likely to be suffering from some level of Alzheimer's Disease. We have also established the minimum number of instances required to reliably achieve this. The experiments were run using various prediction algorithms from the Python Module scikit-learn. We achieved an accuracy level above 55% as measured by stratified k-fold cross validation. The highest score we measured was with a logistic regression classifier, which scored between 55 and 60%. However, other researchers have scored slighter higher using different algorithms and tools to those we used. For instance the research in this Masters thesis [44] claimed to score up to 87%

using similar features on the same NACC dataset with the WEKA j48 classifier. J48 is the name that WEKA uses for the C45 classifier.This score was not produced using a standard validation method in WEKA; rather it was produced by modifying the classifier built by WEKA to optimise the scoring for new instances input to the new classifier in such a way that it scored higher than the 61% they achieved using the WEKA j48 algorithm in its unmodified form. This explained their claim of a higherr score. When they used the WEKA J48 algorithm in its unmodified form their results were closer to ours at 61%. That thesis is one of those listed in the NACC web page of publications using the NACC dataset. They used the WEKA framework for their experiments, but we could not achieve scores at the level they claimed using this framework. We are unsure of the cross validation method used by them in WEKA to calculate their reported accuracy. The WEKA documentation indicates that you can use stratified cross validation, but the thesis does not specifically state that they used it. When we used WEKA with J48 and with logistic regression we got scores of 57.2% and 62.2% respectively, with a 10-fold cross validation in both cases. This is not too far off our scores using scikit-learn and is an indication that our main research question concerning the capability of machine learning to screen possible occurance of Alzheimer's Disease is correct and in line with other research in this field. Another explanation for this difference in results is that they may be using more features from NACC data than we were able to use because we are interested in mapping NACC features to Memory Works clinic data and hence with more features, they are able to get better performance.

Machine learning could also be used in other areas of medical diagnosis. For example [45] reports that an ensemble method called Random Forest was used successfully to assist medical experts in making a diagnosis for both diabetes and breast cancer. Clearly there is a case to be made for research in these and other areas of Healthcare. In fact, scikit-learn comes with a standard diabetes dataset.

We see our contribution in this thesis as reinforcing the concept that machine learning has a place in the futurein helping healthcare practitioners in screening for Alzheimer's Disease, and specifically for the Memory Works clinic dataset. Whilst we would not claim to have produced a model that could be used in a practical, working clinical environment, we feel that such a model is possible. With more research, we feel a model could be developed using existing tools like scikit-learn or WEKA, with better feature selection and other algorithms, as well as a greater number of subject screenings.

The NACC dataset has some autopsy results for a subset of its patients. As autopsy results can provide a definite indication of the presence or absence of Alzheimer's Disease in the patient and the subset could be used by the cross validation methods in scikit-learn to take this fact into account when calculating the prediction accuracy. For example, if a patient has been incorrectly classified as a person with Alzheimer's Disease, but autopsy results prove that this is not the case, then the validation would change the prediction. Scikit-learn is open source, so the cross validation code could be modified to enable this. Alternatively, we could write our own cross validation for use within scikit-learn.

IBM researchers are collaborating on a project with the potential to dramatically increase the accuracy and affordability of early detection through the analysis of voice patterns using some of the same technology found in IBM's Watson system. The idea of screening using voice patterns for machine learning weas found in the literature, so this could be another research topic in which we could utilise these techniques for diagnosis.

The process by which we built and ran our experiments was an iterative one, as described above. At each iteration we added new algorithms. We also added and removed features and tested the effect these changes had on our results. We feel that better results could be attained given different features and algorithms. We have only used a small subset of the algorithms shown in Figure 3.4. Also, with each of the algorithms that we use there are has many hyperparameters which can be modified. We used the default ones in scikit-learn, which are reasonable ones for the majority of learning task.We can not be certain that they are optimum for our training datas. However, Machine learning models in sckit-learn have so called hyperparameters so that their behavior can be tuned for a given problem. Models can have many of these hyperparameters and finding the best combination of parameters can be treated as a search problem. Scikit-learn has a function call Grid search, which is an approach to parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

Doing this might give us better performance in terms of prediction accuracy.

Bibliography

[1] How To Implement Naive Bayes From Scratch in Python: Machine Learning Mindmap. http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/, 2014. Accessed: 15-Jun-2016.

[2] Alzheimer's Association. 2015 Alzheimer's disease facts and figures. *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, 11(3):332, 2015.

[3] Linda Boise, Margaret B Neal, and Jeffrey Kaye. Dementia assessment in primary care: Results from a study in three managed care systems. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 59(6):M621–M626, 2004.

[4] RC Petersen, JC Stevens, M Ganguli, EG Tangalos, JL Cummings, and ST DeKosky. Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review) Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, 56(9):1133–1142, 2001.

[5] Ronald C Petersen, Glenn E Smith, Stephen C Waring, Robert J Ivnik, Eric G Tangalos, and Emre Kokmen. Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, 56(3):303–308, 1999.

[6] Bruno Dubois, Howard H Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T DeKosky, Serge Gauthier, Dennis Selkoe, Randall Bateman, et al. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *The Lancet Neurology*, 13(6):614–629, 2014.

[7] Kate Irving, Paulina Piasek, Sophia Kilcullen, Ann-Marie Coen, and Mary Manning. National Educational Needs Analysis Report. http://dementiaelevator.ie/wp-content/uploads/2013/12/

Elevator-National-Educational-Needs-Analysis-Report-Print-Version.pdf, 2014. Accessed: 2016-06-31.

[8] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. Data mining in Healthcare and Biomedicine: a survey of the literature. *Journal of Medical Systems*, 36(4):2431–2448, 2012.

[9] Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, 7(3):263–269, 2011.

[10] Alzheimer's Disease Neuroimaging Initiatitive: Sharing Alzheimer's Research Data with the World. http://adni.loni.usc.edu/, 2005. Accessed: 17-Mar-2016.

[11] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Jesse Cedarbaum, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, et al. 2014 Update of the Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, 11(6):e1–e120, 2015.

[12] Marilyn S Albert, Steven T DeKosky, Dennis Dickson, Bruno Dubois, Howard H Feldman, Nick C Fox, Anthony Gamst, David M Holtzman, William J Jagust, Ronald C Petersen, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, 7(3): 270–279, 2011.

[13] Olivier Querbes, Florent Aubry, Jérémie Pariente, Jean-Albert Lotterie, Jean-François Démonet, Véronique Duret, Michèle Puel, Isabelle Berry, Jean-Claude Fort, Pierre Celsis, et al. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, 132(8):2036–2047, 2009.

[14] Oskar Hansson, Henrik Zetterberg, Peder Buchhave, Elisabet Londos, Kaj Blennow, and Lennart Minthon. Association between CSF biomarkers and incipient

Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *The Lancet Neurology*, 5(3):228–234, 2006.

[15] Stanisław Adaszewski, Juergen Dukart, Ferath Kherif, Richard Frackowiak, Bogdan Draganski, Alzheimer's Disease Neuroimaging Initiative, et al. How early can we predict Alzheimer's disease using computational anatomy? *Neurobiology of Aging*, 34(12):2815–2826, 2013.

[16] Ramon Casanova, Fang-Chi Hsu, Kaycee M Sink, Stephen R Rapp, Jeff D Williamson, Susan M Resnick, Mark A Espeland, Alzheimer's Disease Neuroimaging Initiative, et al. Alzheimer's disease risk assessment using large-scale machine learning methods. *PloS one*, 8(11):e77949, 2013.

[17] Jieping Ye, Teresa Wu, Jing Li, and Kewei Chen. Machine learning approaches for the neuroimaging study of Alzheimer's Disease. *Computer*, 44(4):99–101, 2011.

[18] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005.

[19] Elizabeth G Kehoe, Jonathan P McNulty, Paul G Mullins, and Arun LW Bokde. Advances in MRI Biomarkers for the Diagnosis of Alzheimer's Disease. *Biomarkers in Medicine*, 8(9):1151–1169, 2014.

[20] Arthur W Toga and Karen L Crawford. The Alzheimer's Disease Neuroimaging Initiative informatics core: A decade in review. *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, 11(7):832–839, 2015.

[21] Kari Antila, Jyrki Lötjönen, Lennart Thurfjell, Jarmo Laine, Marcello Massimini, Daniel Rueckert, Roman A Zubarev, Matej Orešič, Mark van Gils, Jussi Mattila, et al. The PredictAD project: development of novel biomarkers and analysis software for early diagnosis of the Alzheimer's disease. *Interface Focus*, 3(2):20120072, 2013.

[22] de Oliveira A Andrade, Maria Teresa Carthery-Goulart, Júnior PP Oliveira, Daniel Carneiro Carrettiero, and Joao Ricardo Sato. Defining multivariate normative rules for healthy aging using neuroimaging and machine learning: an application to Alzheimer's disease. *Journal of Alzheimer's Disease: JAD*, 43(1):201–212, 2014.

[23] Edward Challis, Peter Hurley, Laura Serra, Marco Bozzali, Seb Oliver, and Mara Cercignani. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage*, 112:232–243, 2015.

[24] Javier Escudero, Emmanuel Ifeachor, John P Zajicek, Colin Green, James Shearer, and Stephen Pearson. Machine Learning-Based Method for Personalized and Cost-Effective Detection of Alzheimer's Disease. *Biomedical Engineering, IEEE Transactions on*, 60(1):164–168, 2013.

[25] Simon F Eskildsen, Pierrick Coupé, Vladimir S Fonov, Jens C Pruessner, D Louis Collins, Alzheimer's Disease Neuroimaging Initiative, et al. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. *Neurobiology of Aging*, 36:S23–S31, 2015.

[26] Saima Farhan, Muhammad Abuzar Fahiem, and Huma Tauseef. An Ensemble-of-Classifiers Based Approach for Early Diagnosis of Alzheimer's Disease: Classification Using Structural Features of Brain Images. *Computational and Mathematical Methods in Medicine*, 2014, 2014.

[27] Chris Hinrichs, Vikas Singh, Lopamudra Mukherjee, Guofan Xu, Moo K Chung, Sterling C Johnson, Alzheimer's Disease Neuroimaging Initiative, et al. Spatially augmented LP-boosting for AD classification with evaluations on the ADNI dataset. *Neuroimage*, 48(1):138–149, 2009.

[28] Elizabeth G Kehoe, Jonathan P McNulty, Paul G Mullins, and Arun LW Bokde. Advances in mri biomarkers for the diagnosis of Alzheimer's disease. *Biomarkers in Medicine*, 8(9):1151–1169, 2014.

[29] L Khedher, J Ramírez, JM Górriz, A Brahim, F Segovia, Alzheimer's Disease Neuroimaging Initiative, et al. Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing*, 151:139–150, 2015.

[30] Stefan Klöppel, Cynthia M Stonnington, Josephine Barnes, Frederick Chen, Carlton Chu, Catriona D Good, Irina Mader, L Anne Mitchell, Ameet C Patel, Catherine C Roberts, et al. Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. *Brain*, 131(11):2969–2974, 2008.

[31] Reisa A Sperling, Paul S Aisen, Laurel A Beckett, David A Bennett, Suzanne Craft, Anne M Fagan, Takeshi Iwatsubo, Clifford R Jack, Jeffrey Kaye, Thomas J Montine, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, 7(3):280–292, 2011.

[32] Christoph Laske, Hamid R Sohrabi, Shaun M Frost, Karmele López-de Ipiña, Peter Garrard, Massimo Buscema, Justin Dauwels, Surjo R Soekadar, Stephan Mueller, Christoph Linnemann, et al. Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, 11(5):561–578, 2015.

[33] Piew Datta, WR Shankle, and Michael Pazzani. Applying machine learning to an Alzheimer's database. In *Artificial Intelligence in Medicine: AAAI-96 Spring Symposium. Stanford*, pages 26–30, 1996.

[34] Sandhya Joshi, V Simha, D Shenoy, KR Venugopal, and LM Patnaik. Classification and treatment of different stages of Alzheimer's disease using various machine learning methods. *International Journal of Bioinformatics Research*, 2(1):44–52, 2010.

[35] William L Jarrold, Bart Peintner, Eric Yeh, Ruth Krasnow, Harold S Javitz, and Gary E Swan. Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease. In *Brain Informatics*, pages 299–307. Springer, 2010.

[36] David A Snowdon, Susan J Kemper, James A Mortimer, Lydia H Greiner, David R Wekstein, and William R Markesbery. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. *Jama*, 275(7):528–532, 1996.

[37] Cati Brown, Tony Snodgrass, Susan J Kemper, Ruth Herman, and Michael A Covington. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2):540–545, 2008.

[38] Karmele López-de Ipiña, Jesus-Bernardino Alonso, Carlos Manuel Travieso, Jordi Solé-Casals, Harkaitz Egiraun, Marcos Faundez-Zanuy, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage, et al. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745, 2013.

[39] Kaye Horley, Amanda Reid, and Denis Burnham. Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech, Language, and Hearing Research*, 53(5):1132–1146, 2010.

[40] Vassilis Baldas, Charalampos Lampiris, Christos Capsalis, and Dimitrios Koutsouris. Early diagnosis of Alzheimer's type dementia using continuous speech recognition. In *Wireless Mobile Communication and Healthcare*, pages 105–110. Springer, 2011.

[41] Calvin Thomas, Vlado Kešelj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Mechatronics and Automation, 2005 IEEE International Conference*, volume 3, pages 1569–1574. IEEE, 2005.

[42] Kay Deckers, Martin PJ Boxtel, Olga JG Schiepers, Marjolein Vugt, Juan Luis Muñoz Sánchez, Kaarin J Anstey, Carol Brayne, Jean-Francois Dartigues, Knut Engedal, Miia Kivipelto, et al. Target risk factors for dementia prevention: a systematic review and Delphi consensus study on the evidence from observational studies. *International Journal of Geriatric Psychiatry*, 30(3):234–246, 2015.

[43] Sandhya Joshi, P Deepa Shenoy, KR Venugopal, and LM Patnaik. Evaluation of different stages of dementia employing neuropsychological and machine learning techniques. In *Advanced Computing, 2009. ICAC 2009. First International Conference on*, pages 154–160. IEEE, 2009.

[44] Henok Wordoffa and Ezedin Wangoria. Alzheimer's Disease Stage Prediction using Machine Learning and Multi Agent System. http://urn.kb.se/resolve?urn=urn:nbn:se:bth-1985, 2012. Accessed: 2016-06-31.

[45] Kemal Polat and Salih Güneş. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4):702–710, 2007.