CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Machine Learning for Predicting Progression of Alzheimer's Disease

Master's thesis in Computer science and engineering

Hildur Egilsdóttir
Hákon Valur Dansson

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2020

Master's thesis 2020

Machine Learning for Predicting Progression of Alzheimer's Disease

Hildur Egilsdóttir
Hákon Valur Dansson

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2020

Machine Learning for Predicting Progression of Alzheimer's Disease

Hildur Egilsdóttir
Hákon Valur Dansson

Supervisor: Alexander Schliep and Fredrik Johansson, Department of Computer Science and Engineering, Erik Portelius, Department of Neuroscience and Physiology at the University of Gothenburg
Examiner: Alexander Schliep, Department of Computer Science and Engineering

Master's Thesis 2020
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Description of the picture on the cover page (if applicable)

Typeset in LaTeX
Gothenburg, Sweden 2020

Machine Learning for Predicting Progression of Alzheimer's Disease

Hildur Egilsdóttir
Hákon Valur Dansson
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

In Alzheimer's disease (AD), amyloid-β ($A\beta$) peptides aggregate in the brain forming plaques. Strangely, these plaques are persistent in both severely cognitively impaired and cognitively normal individuals. Therefore, it is of big value to investigate whether other factors cause some patients with $A\beta$ plaques to have AD dementia and others not. We used data from The Alzheimer's Disease Neuroimaging Initiative (ADNI) to study the differences in individuals with evidence of $A\beta$ plaques and those without. Furthermore, we tried to predict how the cognitive ability of individuals with plaques would progress in the next four years using machine learning techniques. Random forest and elastic net estimators were created, predicting the decline in cognitive test scores as well as diagnosis change of patients only using data from their first visit. The best regression models, predicting the change in cognitive test scores achieved R^2 scores of 0.428 to 0.580 while the classification models, predicting whether a patient will get a worse diagnosis achieved a weighted F_1 score of 0.817. Moreover, patients with $A\beta$ plaques seem to decline faster than those without. The most important features for predicting future cognitive decline were cognitive tests indicating that already cognitive impaired individuals would deteriorate more. Other important factors were fluorodeoxyglucose (FDG) obtained from positron emission tomography and τ proteins measured in cerebrospinal fluid. These models could possibly, with further development, be used in clinical settings as an aid for evaluating how the cognitive function of an individual with $A\beta$ plaques will develop in the near future.

Keywords: computer science, machine learning, Alzheimer's disease, random forest, elastic net, engineering, project, thesis.

Acknowledgements

We would like to thank our supervisors, Alexander Schliep, Erik Portelius and Fredrik Johansson, for their help and guidance in this project. Their great interest in our work and continuous feedback throughout the project was invaluable. Furthermore, we would like to thank Fredrik for always being readily available to answer our questions almost faster than they came up. Thanks also go to The Alzheimer's Neuroimaging Initiative (ADNI) for collecting and providing us with the data used in this project.

Hildur Egilsdóttir and Hákon Valur Dansson, Gothenburg, June 2020

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

Alzheimer's disease (AD) is a serious irreversible disease which can drastically reduce cognitive ability of patients. Memory problems, e.g., forgetting conversations or recent events, are often the first signs of cognitive impairment related to AD. As the disease progresses, such cognitive problems get more severe and eventually, people cannot carry out everyday tasks and become completely dependent on others for their care [27]. According to World Alzheimer's Report 2019 [10], it is estimated that worldwide, over 50 million people have dementia with a new case occurring somewhere in the world every 3 seconds. Furthermore, according to the World Health Organization [28], Alzheimer's may account for 60-70% of dementia cases and is the fourth leading cause of death in developed nations [24]. Because of Alzheimer's lethality and its severe impact on patients' quality of life, a lot of research has been focused on both possible cures and early detection as these are the keys to hindering the progress of the disease.

Even though the disease has been investigated substantially, its high complexity on the pathology level, the genetic and environmental factors involved along with its molecular mechanism leave many questions unanswered. For a long time, it has been believed that the production and deposition of the β-amyloid ($A\beta$) peptide as plaques in the brain is both a disease marker and a partial cause for the cognitive decline of Alzheimer's disease [32]. The levels of $A\beta$ and τ-proteins have been associated with the progression of the disease and their abnormal deposits in the brain are what define AD as a unique neurodegenerative disease [11]. However, several studies have shown that there is only a weak link between these $A\beta$ plaques and the degree of dementia in patients. Furthermore, individuals with plaques don't necessarily exhibit any cognitive impairments [4, 6]. Thus, further studies are needed on the molecular pathogenesis of the disease. One possible explanation could be the existence of patient groups with Alzheimer's disease-like phenotypes but with non-$A\beta$ pathologies contributing to neuronal dysfunction and degeneration. Another could be the existence of factors that shield individuals from developing dementia.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [26] is a project focused on the investigation of Alzheimer's. It is a multicenter study and the data from its database is used in this project. Over 2200 people between the age of 55 and 90 are

divided into groups of cognitively normal elders (CN), elders with mild cognitive impairments (MCI), and elders with dementia. From these participants, ADNI collects clinical data, neuroimaging data, genetic data and biospecimens. The overall goal of ADNI is to validate biomarkers for use in AD clinical treatment trials.

1.2 Aim and purpose

In this project, we use data from ADNI to investigate whether any notable differences in biomarkers, cognitive abilities or other factors can be observed between subjects with $A\beta$ plaques and those without. We further investigate if some features or patterns in subjects with these plaques can explain the differences in severity or absence of dementia observed. To do this, we analyse biomarker data to discover links between them and cognitive decline in subjects and uncover what separates healthy people with AD-like phenotypes from those who suffer from AD. Furthermore, we aim to create a method of progression prediction, i.e., given measurements of a set of key features of a patient, how much, if at all, will his or her cognitive skills decline in the following years. Regression models are built using machine learning methods described in section 2. They predict how much, if at all, a cognitive test score will have declined after either two or four years, using only data from baseline (the first visit of each patient), i.e., the regressors have no knowledge of available future data. Similarly, classification models are built that predict whether people get a worse diagnosis after either two or four years. If the performance of these prediction models will be good, they could be useful for clinicians and could in the future be further developed to assist in diagnosis. Doctors could use them to evaluate a person's risk score for developing AD or how likely, and then how much, patients' cognitive skills would deteriorate in the next few years. If any features are found to be significantly predictive of the progression, they could be targeted for further research, e.g., on the molecular or genetic level.

1.3 Related work

Many researchers have used the ADNI database for analysis of AD patients. For example, Liu et al. [20] used multitask learning to predict the cognitive performance of subjects from MRI data. In their paper, they used two nonlinear multikernel-based multiple learning methods to exploit and investigate a nonlinear relationship between MRI measures and cognitive scores. Another example of a research using ADNI data is from Moradi et al. [24]. They developed a novel technique based on magnetic resonance imaging (MRI) to detect from baseline data whether people with MCI will develop AD within 3 years. First, they presented a new biomarker, utilizing only MRI data, that was based on a semi-supervised learning approach called low-density separation (LDS). Second, they combined this MRI-biomarker with age and cognitive measurements from the baseline to use as features for the learning algorithm. The resulting classifier then predicts whether people with MCI will develop AD within 3 years or not. The results were good and as an example of that, the cross-validated area under the receiver operating characteristic curve

(AUC) score for the classifier was 0.9020. Their main goal was to distinguish between individuals with stable MCI (people who don't develop AD within 3 years) and progressive MCI (people who get AD within 3 years) and thus, cognitively normal people were not taken into account. Additionally, they did not try to find biomarkers which help explain why some people with $A\beta$ plaques get AD-dementia and others not.

We performed explorative analysis, hoping to identify those biomarkers. Therefore, we only looked at amyloid positive people (those who have strong indicators of having $A\beta$ plaques) at baseline, both CN and MCI elders, and tried to predict their cognitive decline in two or four years. By doing this, we hoped to find the aforementioned distinguishing factors. While Moradi et al. [24] used MRI images, cognitive measurements and age, and Liu et al. only used MRI data, we included almost all available data including data extracted from cerebrospinal fluid (CSF), positron emission tomography (PET) scans, plasma and serum. On the other hand, we did not look at any raw images, neither MRI nor PET scans.

1.4 Limitations

The data in this project are limited to the available data from ADNI during the time the project is conducted and no other data are considered. More measurements and features are continuously being added to the data which might fill in some of the missing observations but we only use data that were available on 27 February 2020. The progression predictions are limited to $A\beta$-positive subjects at baseline. First, this limits the data to the 1210 subjects with both $A\beta_{40}$ and $A\beta_{42}$ measurements and then further limits the data to the $A\beta$-positive cohort resulting in 681 subjects. Only numerical and categorical features are used though more data are available. Therefore, a lot of available and unprocessed data are not used such as raw images, genome sequences and physical fluid samples.

1.5 Ethics

Ethical challenges include the risks of compromising the privacy of subjects. The data used in this project have been anonymised. The names or other identifying factors for patients have been removed and instead, they are identified by roster identifiers in the data. However, it might be possible to trace the data back to specific subjects since, for example, their genomes are stored in the database. Considering that many of the subjects are in a difficult situation needing expensive medical care and are likely to have mental and/or cognitive impairments, they might be more easily manipulated than healthy individuals. A data breach might, therefore, result in malicious people taking advantage of them. Thus, we only store data and code on locked devices and in private repositories. Raw data were not shared with anyone. As this is a very difficult disease that rapidly degrades the patients' quality of life, one has to be careful not to push potential findings too hastily for personal gain, potentially giving false hope or deceiving those in need.

2

Background

2.1 Alzheimer's disease diagnosis

Alzheimer's disease was initially defined as a clinical-pathological entity. The only possible way to diagnose a person as definitely having the disease was if a patient had the clinical criteria for probable Alzheimer's disease as well as histopathologic evidence obtained from biopsy or autopsy. In living patients, the diagnoses were only probable or possible AD [22]. A research framework proposed in 2018 by the National Institute on Aging and Alzheimer's Association (NIA-AA) [11] instead defines AD in vivo by biomarkers and by postmortem examination, not by clinical symptoms. In the research framework, a classification system is proposed called AT(N) where the biomarkers are split into 3 classes, $A\beta$ deposition (A), pathologic τ (T), and neurodegeneration (N). $A\beta$ deposition biomarkers are based on PET or CSF where either $A\beta_{42}$ or the $A\beta_{42/40}$ ratio are considered valid indicators [3]. Pathologic τ may be measured in CSF and through PET scans while neurodegeneration can be estimated by biomarkers from FDG PET, MRI or total-τ in CSF [11].

The abnormal $A\beta$ deposition is what determines if individuals have the Alzheimer's pathological change and they are said to have AD if it is combined with abnormal pathologic tau. Table 2.1 shows how the subject is classified with regards to Alzheimer's using binary classes in the AT(N) model [11]. It may be that $A\beta$ and τ levels are not the *cause* of disease development but their abnormal deposits in the brain uniquely define AD.

2.2 Pre-processing

2.2.1 Data standardisation

The ADNI data include features of different units and scales. This can be problematic in many machine learning algorithms as some techniques assume, for example, that the data are distributed around 0 with uniform variance. When one feature has a much higher variance, it may dominate the objective function of a machine learning model, such as the ones described in subsection 2.3.2, and as a result, the learning ability of the model can be hindered significantly. To avoid this problem, it can be beneficial to normalise the data [31]. Furthermore, it is a requirement for

Table 2.1: The table shows how the Alzheimer's profiles are defined in [11] by binarising the three AT(N) biomarker types as either positive(+) or negative(-).

AT(N) profiles	**Biomarker category**	
$A^-T^-(N)^-$	Normal AD biomarkers	
$A^+T^-(N)^-$	Alzheimer's pathologic change	Alzheimer's continuum
$A^+T^+(N)^-$	Alzheimer's disease	
$A^+T^+(N)^+$	Alzheimer's disease	
$A^+T^-(N)^+$	Alzheimer's and concomitant suspected non Alzheimer's pathologic change	
$A^-T^+(N)^-$	Non-AD pathologic change	
$A^-T^-(N)^+$	Non-AD pathologic change	
$A^-T^+(N)^+$	Non-AD pathologic change	

some machine learning methods. Normalising does not affect tree-based algorithms, however. This is because, at each point in the tree, only a single feature is being used for splitting. Therefore, the scale of one feature does not affect the evaluation of another feature. Standardisation (also known as Z-score normalisation) is the method used in this project:

$$x_i' = \frac{x_i - \mu_i}{\sigma_i} \tag{2.1}$$

In this formula, the mean μ_i and standard deviation σ_i are found for each feature i individually. Then, each value x_i is updated to the standardised value x_i'. After this transformation, each feature has a mean of 0 and a standard deviation of 1. Thus, features on large scales are less likely to dominate the objective function.

2.2.2 One-hot encoding

The data contain a mixture of continuous and categorical features. For most machine learning models a numerical representation of features is needed for them to work. This can be solved by coding categorical classes as different numbers. However, for linear models and tree-based models, this implies a relationship between the different categories that may not be present, e.g., if the classes cat, dog and horse are coded as 1, 2 and 3 respectively, a linear model considers dogs and horses more similar than cats and horses. Instead, one-hot encoding is often used. With this approach, categorical values can be represented with multiple binary features with value 1 if the data point is a member of a specific class, and 0 otherwise. One-hot encoded features should not be standardised like continuous variables as they are already on a similar scale (0/1). If prevalence is far from 50%, standardising will introduce a large factor into their relative transformed scales.

2.2.3 Dealing with missing data

The data set used has a lot of missing values, i.e., each data point does not necessarily include values for every feature. Since most machine learning models, including the ones used in this thesis, cannot deal with missing data, this problem needs to be addressed.

One solution is to exclude all incomplete observations by removing the rows with any missing values. By using this method, only a subset of the available data would be considered and the size of the patient cohort would, consequently, be reduced substantially. Therefore, this method causes loss of information, efficiency and predictive power. Furthermore, it can potentially lead to serious biases if the missing observations are not completely at random due to differences between the observed and the missing data [2, 13]

Another way of dealing with this problem is by using imputation, i.e., replacing the missing values with other plausible values, estimated using available information. The model can then use the resulting values as the observed ones and ignoring of missing data can thus be avoided. For this method to be useful, the distributional relationship between the available and missing data needs to be captured, which can be a difficult task. Furthermore, it is necessary to keep in mind that the imputed values are not the real values and the resulting predictions all come with uncertainty.

Additionally, it can be the case that the gaps in the data are not completely at random and the knowledge that a particular observation is missing may contribute to the output of the models. To preserve and capture the correlation between missing data and the output, binary features can be added, indicating whether an observation was originally absent or not. Such correlations are common in medical data and when the absence does predict the outcome, the use of a missing indicator can be a good approach [35]. The inclusion of an indicator must be combined with an imputation method for the missing observations. This allows each participant to be included in the analysis to maintain statistical power.

2.3 Machine learning algorithms used

2.3.1 Random forest

Random forest [5] is a popular machine learning algorithm which is flexible with a wide range of applications and can model complex patterns and relationships. Random forests are classifiers built on decision trees and can be expanded to regression tasks. A decision tree is a tree-like data structure as shown in figure 2.1 where its internal nodes represent decisions about the data, splitting the data set into two or more subsets. The branches from these nodes represent outcomes of the decision, and each leaf in the tree represents a class label. Such a tree represents a recursive algorithm that uses input variables to predict the target class of each data point. Training of decision trees is typically done by finding a feature that splits the data into two (or more) subsets that are more homogeneous than the original data set.

Then, for each of the resulting data sets, new decision trees are trained recursively until one or several leaves are reached. A few different methods can be used to split the data at each node in the tree.

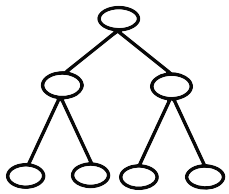

Figure 2.1: An example of a tree structure.

In this project, we used the gini impurity method. It can be described as follows [36]: A training set T is given and $p_i = p(i|T)$ is the proportion of examples in T belonging to class i where $i \in \{1, 2, \ldots, k\}$ is the class label. Then the gini impurity is defined as follows:

$$I(T) = \sum_{i=1}^{k} p_i(1 - p_i) \tag{2.2}$$

The gini impurity measures the probability of misclassifying a randomly picked element by randomly classifying it according to the class distribution in the data set, i.e., the lower the value, the better. The splitting operation splits the data set into two subsets T_1 and T_2. So to find the best position of a split we maximise the following function called a gini gain:

$$I_{total}(T) = I(T) - I(T_1)p(T_1) - I(T_2)p(T_2) \tag{2.3}$$

where $p(T_1)$ and $p(T_2)$ denote the proportion of data points going to the left and right subtrees, respectively.

A single decision tree tends to overfit to the training data. To avoid that, random forest classifiers consist of several decision trees. Each tree predicts an output class for the input and the most frequent class becomes the forest's output. This results in less overfitting and it has been proved that the generalisation error converges as the number of trees in a random forest increases [5].

For random forests to work as well as possible, the decision trees, and their predictions, should be uncorrelated. This can be achieved by using a method called bagging. When using this method, having a training set T of size N, we create m new training sets $T_1, T_2, \ldots, T_m$, each of size N by randomly sampling with replacement from T. A decision tree is then trained separately on each of the training sets, T_i. The output of the random forest classifier is the mode of all the trees. Moreover, when splitting each node in the tree, the best split is either found using all of the features or a random subset of the features available.

Random forest regression models work in the same way as the classifiers. However, instead of each leaf in the decision trees representing a class label, it represents an

output value prediction. The random forest then outputs the mean prediction value of the trees in the forest.

2.3.2 Linear models

Linear regression models a linear relationship between one or more dependent variables y and one or more explanatory variables x. The model prediction function for a single dependent variable y and explanatory variables x_i, $i \in \{1, 2, \ldots, p\}$ can be written as $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\beta}) = \beta_0 + x_1\beta_1 + ... + x_p\beta_p$ where $\boldsymbol{\beta}$ are unknown parameters or coefficients of the linear model. The relationships are modelled by estimating the unknown parameters from the data. This estimation is done by minimising a cost function based on how well the data fit to the proposed linear relationship. The cost functions used to build our linear regression models are:

- **Least squares** finds the optimal function by minimising the sum, S, of squared residuals:

$$S = \sum_{j=1}^{N} (y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta}))^2 \tag{2.4}$$

  Here, N is the number of observations in the data. With standardised data the minimisation can be written compactly in matrix form as:

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \right\} \tag{2.5}$$

- **Ridge regression** introduces, on top of the least squares, a so called L_2 norm regularisation term that puts the constraint $\sum \beta_i^2 = c$ on the $\boldsymbol{\beta}$ coefficients. When introducing such a constraint, it is necessary to have standardised the explanatory variables $\boldsymbol{x}$ or the shrinking of the coefficients caused by the regularisation will mostly depend on the magnitude scale of the inputs, not their actual contribution to the model. When the predictor variables are highly correlated, ridge regression produces coefficients which predict and extrapolate better than least squares [21]. By using the optimisation method of Lagrange multipliers the minimisation subject to the constraint can be written as:

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} \tag{2.6}$$

  The λ is a selected parameter that controls the amount of regression and the exact relationship between c and λ is data-dependent.

- **Lasso regression** uses an L_1 norm regularisation term with the least squares. The L_1 norm introduces the constraint $\sum |\beta_i| = t$. Comparing to the ridge regression, which only shrinks the coefficients, the lasso also reduces coefficients, that do not substantially contribute, to zero [33]. As with the ridge regression, the input needs to be standardised. The minimisation is formulated as:

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \tag{2.7}$$

- **Elastic net** combines both the L_1 and L_2 regularisation terms. It has similar sparsity of representation as the lasso while also encouraging the ridge grouping effect of correlated predictors. The elastic net has been shown to be particularly useful when the number of predictors p is large and the observations n are relatively few as it is in our case [38]. The minimisation problem of elastic net can be written as:

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\} \tag{2.8}$$

2.3.3 Logistic regression

logistic regression uses the logistic function which is a sigmoid function, that takes any real input t and outputs a value between zero and one. In binary classification, the output of the logistic function can thus be interpreted as the probability of being assigned the class 1. The input t can then be set as an underlying linear model dependent on explanatory variables $\boldsymbol{x}$ and linear coefficients $\boldsymbol{\beta}$.

$$t = f(\boldsymbol{x}, \boldsymbol{\beta}) = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p \tag{2.9}$$

This gives the probability of being assigned to class 1 as:

$$P(1|\boldsymbol{x}) = \frac{e^{f(\boldsymbol{x},\boldsymbol{\beta})}}{e^{f(\boldsymbol{x},\boldsymbol{\beta})} + 1} \tag{2.10}$$

When estimating the linear coefficients $\boldsymbol{\beta}$ of the model they can, as in linear regression, be subject to the same regularisation constraints such as the L_1 and L_2 norms in elastic net. However, unlike linear regression, it is not possible to find a closed-form expression for the coefficient values that maximise the likelihood function, so an iterative method such as gradient descent must be used instead.

2.4 Feature selection

The final data set used in this project has 864 features and 2253 subjects. Since high-dimensional data can have a high degree of irrelevant and redundant information which can greatly degrade the performance of learning algorithms [37], a method for feature selection is needed. The problem of feature selection is defined as follows: given a set of d features, select a subset of size m that leads to the smallest classification error [12]. In some cases, prior knowledge of the relationships between features and outputs is enough and the features can thus be handpicked. In other cases, the use of algorithms is necessary for the task. An exhaustive search for the optimal set of features would be infeasible for so many features since it is known to be NP-hard [1]. On the other hand, greedy search strategies are considerably faster even though they may not find the optimal set of features.

2.4.1 Forward selection

An example of a greedy search strategy is the sequential forward selection strategy. When using this strategy, we start with an empty set of features and select the single best feature, i.e., the feature that gives the best performance for the model. When a feature has been selected, the next feature added is the one that gives the best model performance along with the already chosen feature. Thus, once a feature has been chosen, it will not be discarded from the model. This step of selecting and adding a feature is done until a stopping criterion is reached. This criterion can, for example, be that no feature improves the model performance for a predefined number of rounds or that a maximum number of features has been added to the model.

2.5 Cross-validation

Cross-validation is a method to evaluate how well a machine learning model generalises, i.e., how it will perform on unseen data. In k-fold cross-validation, a data set D is split into k mutually exclusive random samples $D_1, D_2, \ldots, D_k$ of approximately the same size. Then, for each of these subsets D_i where $i \in \{1, 2, \ldots k\}$, the model is trained on $D \setminus D_i$ and then validated on D_i. This is repeated so that each subset is used exactly once as a validation set. The overall performance can then be estimated by averaging over the k performances. In stratified cross-validation, the splits are done so that each sample contains approximately the same distribution of classes as the original data set [14].

Other generalisation performance estimators include leave-one-out and hold-out. In the former, a data set of size n is trained on (n-1) samples and evaluated on the single remaining sample. This is done for each sample in the training set. In the latter, the data are split into two parts, one for training and the other for testing. The k-fold cross-validation estimate has a lower bias than the hold-out method and is cheaper to implement than the leave-one-out method [12].

2.6 Model performance evaluation

2.6.1 Classification models

For the classification models, we used the F_1 weighted score to evaluate their performance. The formula for F_1 score is

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.11}$$

where

$$recall = \frac{true\,positive}{true\,positive + false\,negative} \tag{2.12}$$

and

$$precision = \frac{true\,positive}{true\,positive + false\,positive} \tag{2.13}$$

are calculated using the confusion matrix of the classifier as shown in figure 2.2. Using a weighted F_1 score means that the F_1 score is calculated for each variable and averaged using weights dependent on the number of true labels for each class. The possible range of the F_1 score is from 0 to 1, the higher the value, the better the model is considered.

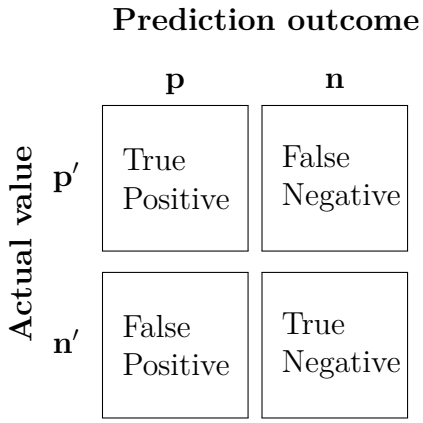

Figure 2.2: Confusion matrix of a binary classification model where p and n stand for positive and negative predictions respectively and p' and n' stand for actual values being positive and negative respectively.

2.6.2 Regression models

For regression models, we use the coefficient of determination, denoted by R^2. This method represents the proportion of variance in the output variable that is explained by the input variables. If $\hat{y}_i$ is the model prediction for the real value y_i of variable $i \in \{1, 2, \ldots, N\}$, the estimated R^2 is defined as

$$R^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2.14}$$

where $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$.

If the modelled values are exactly the same as the real values in every case, the R^2 score of the model would be 1. A model always predicting the expected value of y regardless of the input variables would have an R^2 score of 0. Furthermore, models can have a negative R^2 score if they perform worse than a mean-predicting model.

3

Data

All of the data used in this project come from ADNI, a multicenter study which collects data and stores samples for research. The data contain 2253 unique subjects and statistics for features frequently mentioned in literature on Alzheimer's can be seen in table 3.1. A considerable amount of the data is measured by different labs with samples collected by ADNI, such as blood or CSF fluid. These labs often conduct specific research and usually, measurements are done for only a subset of the subjects. This means that not all data are available for every subject and that some measurements may have been calibrated differently. Moreover, the data and specimen are collected from three consecutive studies made by ADNI, each after receiving additional funding [34]. These different origins of data result in some features having little to no overlap with other features, i.e, no or few subjects have both feature A and feature B. This problem of lacking overlap and difference in feature availability can be seen in figure 3.1 in the features AV45 and PLASMATAU, which, even though respectively, 1050 and 575 subjects have these measurements, no subject has both.

Table 3.1: Baseline cohort statistics.

		Missing	Overall	CN	MCI	Dementia
n			2253	815	1020	391
Age, mean (std)		3	73.2 (7.2)	72.9 (6.2)	72.9 (7.6)	74.8 (7.9)
Gender, n (%)	M	0	1195 (53.0)	360 (44.2)	603 (59.1)	220 (56.3)
	F		1058 (47.0)	455 (55.8)	417 (40.9)	171 (43.7)
MMSE, mean (std)		0	27.4 (2.7)	29.1 (1.1)	27.6 (1.8)	23.2 (2.2)
ADAS13, mean (std)		24	16.9 (9.3)	10.4 (4.6)	17.0 (6.7)	30.2 (8.0)
TAU, mean (std)		1038	287.0 (132.8)	237.5 (90.6)	287.1 (134.8)	366.6 (145.8)
$A\beta_{42}$, mean (std)		1043	1094.7 (610.3)	1373.2 (644.7)	1056.4 (576.0)	745.6 (404.8)
FDG, mean (std)		799	1.2 (0.2)	1.3 (0.1)	1.2 (0.1)	1.1 (0.1)
APOE4, n (%)	0	202	1116 (54.4)	529 (69.7)	466 (50.0)	119 (33.3)
	1		741 (36.1)	209 (27.5)	364 (39.1)	167 (46.8)
	2		194 (9.5)	21 (2.8)	102 (10.9)	71 (19.9)
Hippocampus, mean (std)		761	6790.1 (1185.3)	7396.2 (909.4)	6778.3 (1132.9)	5761.9 (1023.2)
AV45, mean (std)		1203	1.2 (0.2)	1.1 (0.2)	1.2 (0.2)	1.4 (0.2)
$A\beta$-ratio, mean (std)		1043	0.133 (0.056)	0.160 (0.054)	0.130 (0.056)	0.098 (0.035)
$A\beta$-positive, n (%)	0	1043	529 (43.7)	248 (67.6)	248 (40.1)	33 (14.7)
	1		681 (56.3)	119 (32.4)	370 (59.9)	192 (85.3)

It can be seen in table 3.2 that the data are far from complete since for every patient only a subset of the discussed data types is available. Thus, to fully utilise the data, an imputation strategy is needed for dealing with missing data. Furthermore, patients can continue in the study and come for follow up checks at fixed intervals generating a sequence of data-points for each patient. However, the number of follow-ups for patients varies as well as which measurements are done. Moreover, the number of observations generally decrease with time. Comparing figures 3.1 and 3.2 shows the decrease or total lack of observations for features two years after baseline.

Table 3.2: Information about the number of patients where specific data are available for each visit. The columns are the number of months since baseline (3m = 3 months after baseline inspection).

	Baseline	3m	6m	12m	24m	36m	48m	60m	72m
Unique subjects	2253	793	1618	1793	1370	853	716	478	508
Diagnosis	2226	0	1616	1618	1337	821	687	427	449
Age	2250	793	1618	1792	1370	853	716	478	508
Aβ_{42}	1210	0	1	67	247	40	174	38	21
TAU	1215	0	2	320	443	81	183	43	28
MMSE	2253	0	1614	1620	1336	816	688	424	449
ADAS13	2229	0	1596	1604	1308	799	681	420	446
All of the above	1193	0	1	66	245	40	161	37	21

3.1 Target values for machine learning models

A lot of clinical data are available such as cognitive tests and disease state diagnoses. We include the cognitive tests *The Alzheimer's Disease Assessment Scale Cognitive Subscale - 13 items* (ADAS13) and *Mini Mental State Examination* (MMSE) as regression targets. ADAS13 is a scale from 0 to 85 where a higher score means worse cognitive function. It includes 13 different tasks: word recall, naming objects and fingers, commands, constructional praxis, ideational praxis, orientation, word recognition, language, comprehension of spoken language, word-finding difficulty, remembering test instructions, delayed word recall and number cancellation or maze task [15]. MMSE, on the other hand, is a scale from 0 to 30 where lower score means worse cognitive function. It includes tests of the following: orientation to time, orientation to place, registration, attention and calculation, recall, language, repetition and complex commands [7]. We use the clinical dementia diagnosis as part of the classification class for the machine learning algorithms. The diagnoses are split into the three categories mentioned in section 1.2: dementia, MCI and CN.

3.2 Features used

The data contain analysed biofluid samples from CSF, plasma and serum for which we included measurements of different biochemicals such as proteins, hormones and lipids. Medical imaging data, such as PET scan and MRI data are available for many patients. We only use numbers extracted from analysed images, i.e., the images are only used indirectly and not the raw images. Furthermore, patient data such as age, education and gender are included. Genetic data are available but we only include the APOE4 gene which is the main genetic risk factor for AD [19]. A person can have zero, one or two copies of the gene.

The CSF data include both τ and $A\beta$ which can be used to assess if their levels are abnormal in the brain since that is what defines the disease as mentioned earlier. The CSF $A\beta$ data include measurements of both $A\beta_{42}$ and $A\beta_{40}$, which are $A\beta$ peptides ending at positions 42 and 40 respectively. Their ratio in CSF measurements has been proposed to better reflect brain amyloid production than their individual measures [17, 18]. Therefore, the ratio $A\beta_{42}/A\beta_{40}$ ($A\beta$-ratio) in CSF is calculated and added as a new feature for all subjects with both measurements available.

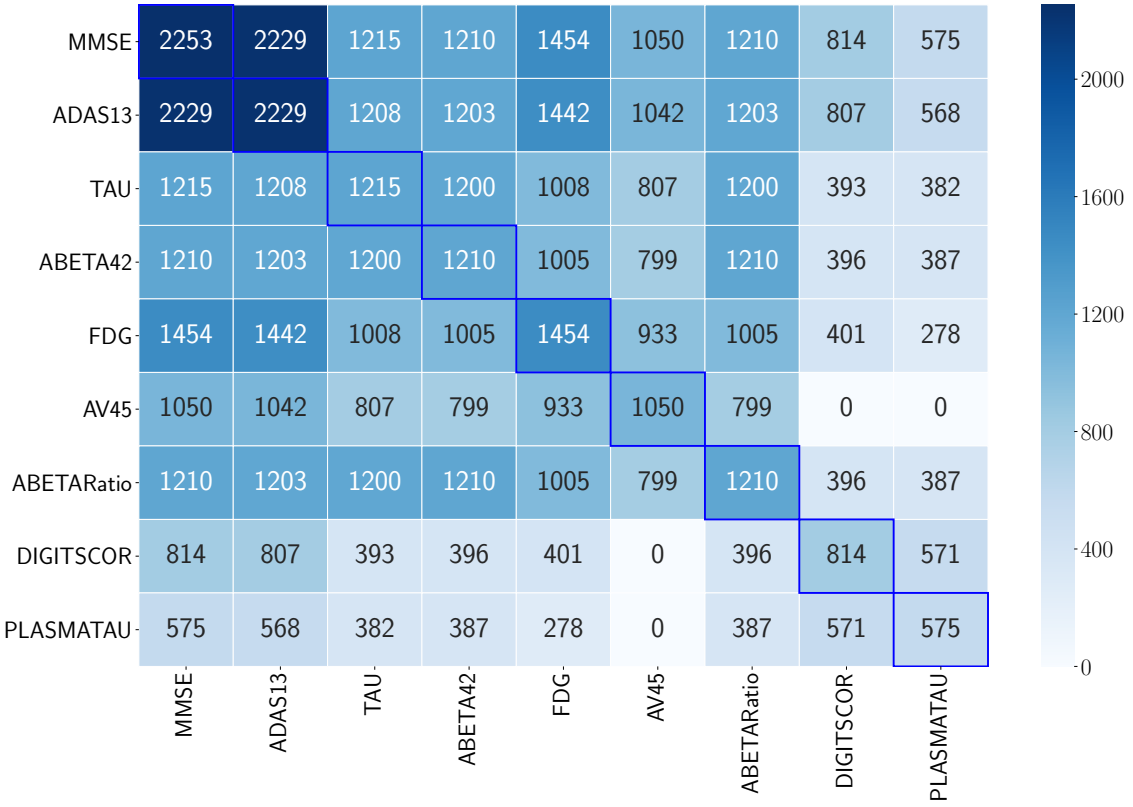

Figure 3.1: Matrix showing how many subjects have measurements for certain pairs of features at baseline. As can be seen from, for example, the AV45 and PLASMATAU, some features do not have any subjects with both measurements. It also shows there is a large difference in the number of subjects each feature was measured for.

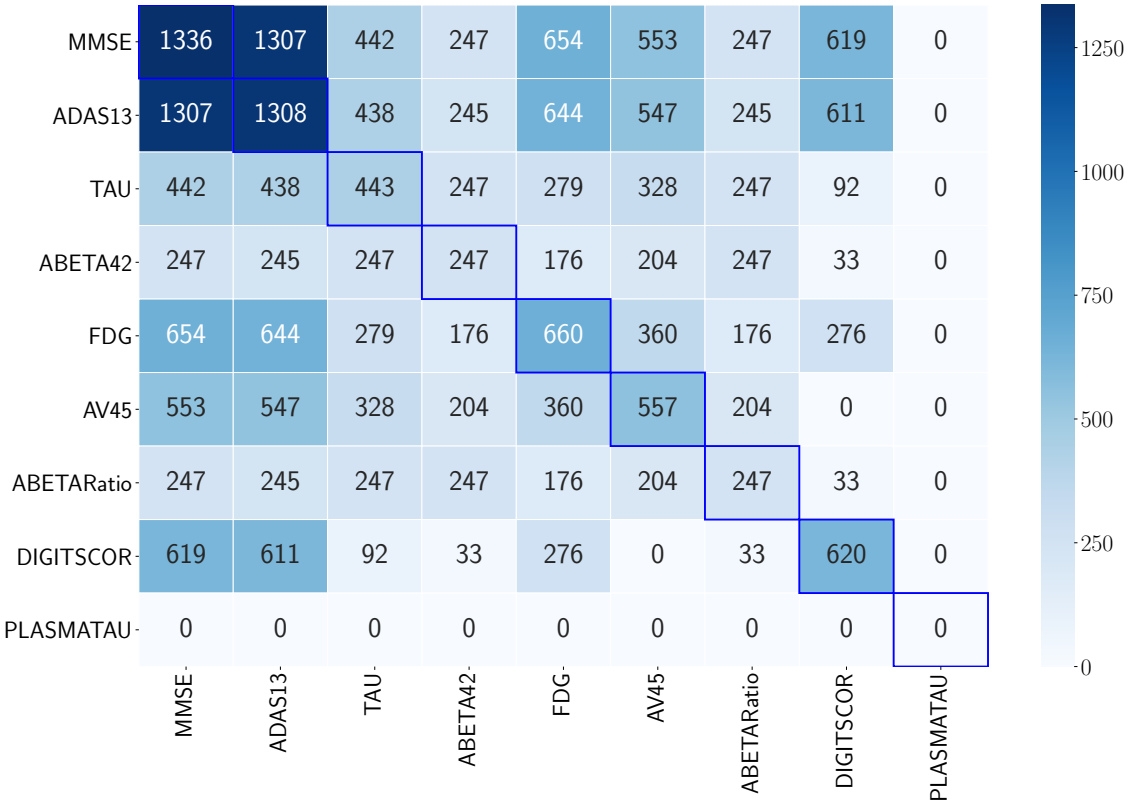

Figure 3.2: Matrix showing how many subjects have measurements for certain pairs of features two years from baseline. As can be seen by comparing to figure 3.1, the number of measurements has decreased and PLASMATAU is not observed for any subjects.

4

Methods

4.1 Data construction

The data from ADNI are split between many folders and files. We started by using one file (described as key variables merged into one file) as a base file and then, we added the other files including biospecimen data. This was done by joining the files on the participant roster id (RID), which is used to identify each participant in the study and the visit code, which indicates from which point in time of the study the information originates. The visit code at the first visit of each subject is called baseline and each subsequent visit of that subject is denoted by the number of months since baseline. Eventually, we had 14570 lines and 1052 features in the data frame created from 2253 unique participants. Each line represents data for one patient at a single visit. Some features were irrelevant to the project or contained very few measurements and were therefore removed from the data set. 188 such features were removed, e.g., comments and sample identifications. The total imported features were thus 864 before adding one-hot encoding and missing indicators.

When the data were assembled, we started by visualising availability. We then went on to explore the differences between subjects based on their diagnosis. Furthermore, we visualised the diagnoses by creating scatter plots of key features such as $A\beta$-ratio vs. p-τ with colours denoting different diagnoses. The resulting graph can be seen in figure 5.1.

4.2 Differences based on $A\beta$-ratio

Based on the measure of the $A\beta$-ratio in CSF at baseline, subjects were defined to have $A\beta$ pathology or not. They were divided into two groups, those who had a lower ratio than 0.13 ($A\beta$-positive) and those who had a higher ratio ($A\beta$-negative). This particular split value was based on the observation that the data resemble two normal distributions which could be fairly well separated at a slightly higher ratio. Furthermore, by visual inspection of figure 4.1, the value 0.13 was chosen. The split can also be seen in the histograms in figure 4.2. The chosen split has a higher ratio than the 0.0975 proposed in [17] for diagnosis of AD. This is reasonable since we were also interested in subjects who have not developed AD dementia but may have $A\beta$ plaques. The data for each of the different subject diagnosis groups, CN, MCI and

dementia were then investigated further based on the two $A\beta$ groups. This was done by performing statistical analysis and visualising the different cognitive decline from plots of the MMSE test scores over time shown in figure 5.3. Furthermore, violin plots (figure 5.2) and other graphs were made to evaluate the differences between the groups. This exploration was done before imputation to avoid viewing possible non-existing patterns generated by the imputation.

Next, we wanted to predict how patients with $A\beta$ pathology would progress and explain why there is such a large variation in cognitive abilities of subjects with the pathology. Using only the $A\beta$-positive group, classification and regression models were built for this prediction. All models were built using scikit-learn [29], a machine learning library that implements the algorithms described in chapter 2.

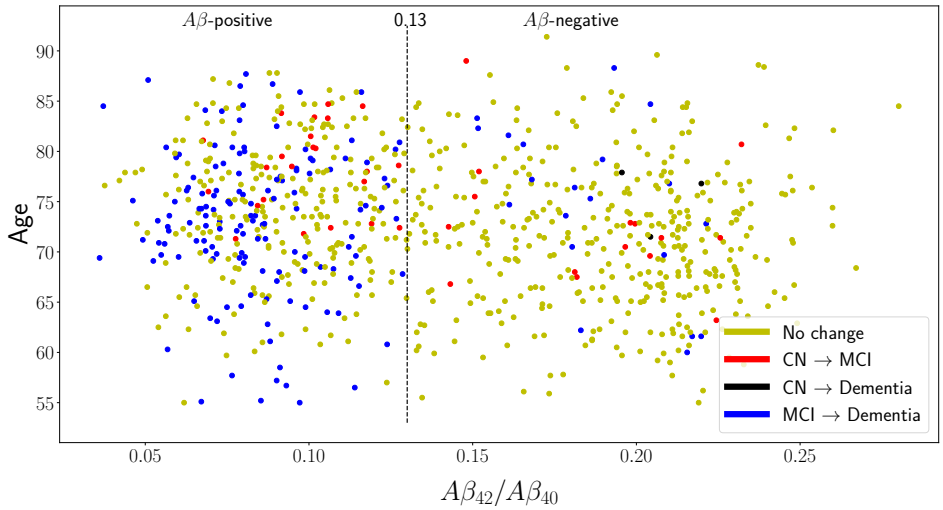

Figure 4.1: Scatter plot showing the $A\beta$ ratio versus the age of the subjects at baseline. The different colors represent how the diagnoses change after two or four years. The black dotted line indicates the split value for the ratio, 0.13.

4.3 Classification

Target values for the progression of dementia were created from measurements of existing features at different times. For classification, a binary variable was created, indicating whether or not a subject's diagnosis had worsened in two or four years. Some subjects only had data for four years after baseline while others only had data from two years. To maximise the amount of available data for the classifiers, the diagnosis-change feature combined both and indicated whether a person had a worse diagnosis after either two or four years.

Using only the $A\beta$-positive subjects, classification models were built to classify whether or not subjects would deteriorate. These models predicted if a subject would go from being in the CN group at baseline and convert to either MCI or dementia, or go from MCI at baseline to dementia after either two or four years. Two

types of classification models were built, a random forest classifier and an elastic net logistic regression classifier. The target values of the models were not imputed nor the $A\beta$-ratio. This was done so that we would not try to model some non-existing patterns or include people who were not really $A\beta$-positive. The number of $A\beta$-positive subjects including a diagnosis after either two or four years was 681.

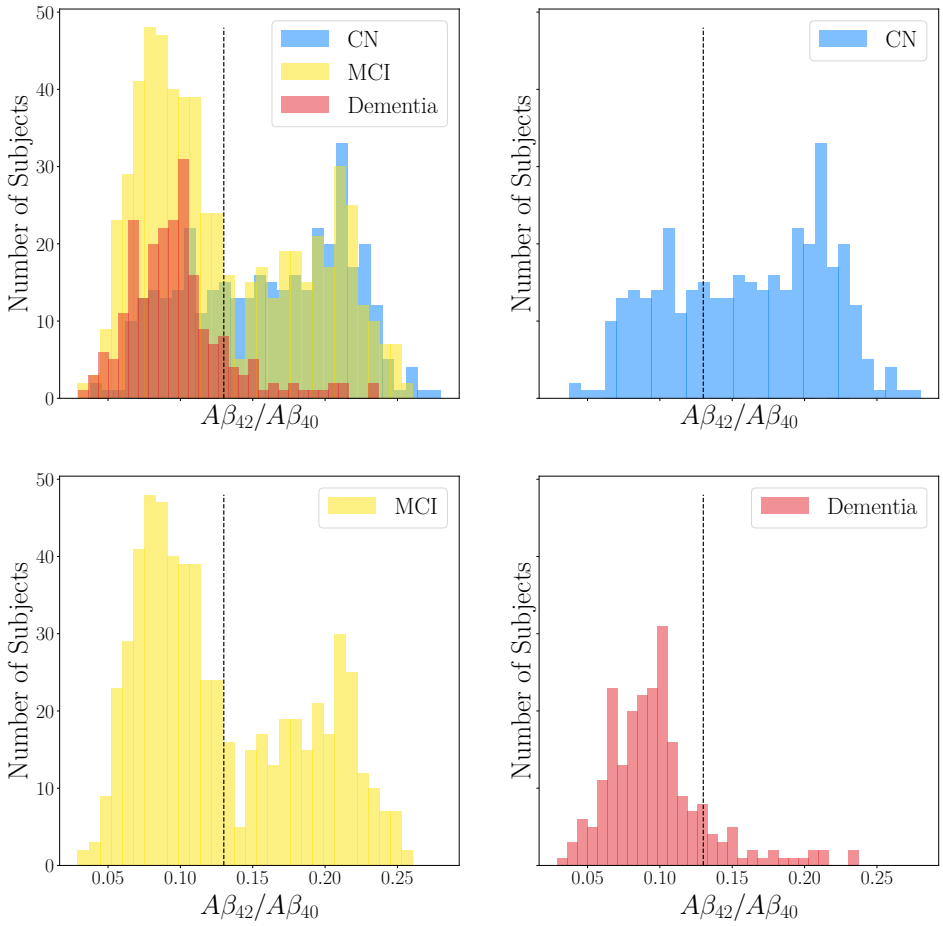

Figure 4.2: Histogram showing the $A\beta$ ratio of the subjects at baseline. The different color-groups represent the different diagnoses: dementia, MCI and CN. The upper left histogram shows all diagnoses together and overlaps have combined colors like green and dark red.

4.4 Regression

We added the values for change in ADAS13 score after two and four years and the change in MMSE score after two and four years to use them as regression target

values. As for the classification task, only the $A\beta$-positive subjects were included in the regression task.

Four different estimators were built (ADAS13-score change after two years, ADAS13-score change after four years, MMSE-score change after two years and MMSE-score change after four years) and for each of them, both a random forest and an elastic net model were created. The number of $A\beta$-positive subjects with cognitive scores after two years was 483 and 227 after four years.

4.5 Preprocessing

After combining all the data into a single data set, a missingness indicator was added for each feature, i.e., 1 if the feature was originally missing and 0 if the feature was available. Next, the features were defined as either continuous or categorical and the two groups were processed in different ways. Categorical values were one-hot encoded and the imputed values were set to 0. For continuous features, missing values were imputed using the mean value of the feature before standardising. This way, the amount of usable data was maximised and all features were on a similar magnitude scale.

4.6 Feature selection

All features could be used in both elastic net and random forest as they both favour features that substantially contribute as predictors. Thus, estimators of both types were trained using all features. However, they might perform better if only a selection of useful features were used. Therefore, several feature selection methods were evaluated.

First, for each feature in the data, both types of estimators were trained to determine the estimators' performance using a single feature. Second, the forward selection was done by adding to the selection, at each iteration, the feature that improves the average weighted F_1 score (classification) or R^2 score (regression) the most. Due to computation time, hyperparameter selection was not included during the forward selection. New features were added to the selection until the last two features added had not improved the score of the model or if it reached a maximum of 25 additions. When adding features, a continuous feature and its indicator were considered a pair and evaluated and added together. Similarly, all the one-hot encoded classes of a categorical value were considered a group and evaluated and added together to the selection. This means that if the maximum of 25 additions was done at least 50 features were included. Forward selection was performed for both estimator types. Finally, some hand-picked selections were done based on the features that performed well in the previously mentioned selection methods. This selection included all features selected by at least two models using forward feature selection.

4.7 Evaluation

The classification models were scored using the weighted F_1 score while the regression models used the R^2 score as a criterion. To get a generalised evaluation, 5-fold cross-validation was used: during each validation run the estimator was trained on 80% of the data and tested on the other 20%. The average test score was then given as the final score for the model. Further evaluation was done by inspecting the confusion matrices of the binary prediction. The regression models were further evaluated by visualising the true values versus the predicted values, in so-called calibration plots. The feature importance was also visualised by plotting the 15 features that had the highest importance values. Feature importance was used as a term for random forest feature importance and the absolute values of the coefficients for elastic net.

4.8 Hyperparameter selection

To select the best hyperparameters, a search was performed during the training of a model. The previously split training data (80%) was further split into 5-folds (stratified for classification). Then, for each fold, a grid search was performed by training the model on the other four folds, resulting in a 5-fold cross-validated hyperparameter search. The hyperparameter values included in the grid search can be found in tables A.1 and A.2 in appendix A. These models were then retrained with the selected hyperparameters using the full training split (80%) and tested on the test set, i.e., each model was trained on the same part of the data that was used to select its hyperparameters and tested on the rest. This gave five models for each estimator, each of them possibly with different hyperparameters and the final model performance was then the average performance over these five resulting models.

5

Results

5.1 Comparing $A\beta$-positive and negative groups

A scatter plot of the $A\beta$ ratio versus PTAU created from baseline data and colored by diagnosis is shown in figure 5.1. When looking at the scatter plot, it is evident that $A\beta$-negative patients, i.e., those with a higher $A\beta$ ratio than 0.13, are not likely to have dementia at baseline. That side of the scatter is predominantly blue and yellow, indicating CN and MCI subjects respectively. On the other hand, the subjects in the upper left corner of the image, being $A\beta$-positive and having high PTAU values seem to have worse diagnoses. Furthermore, MCI patients seem to be spread over the whole spectrum.

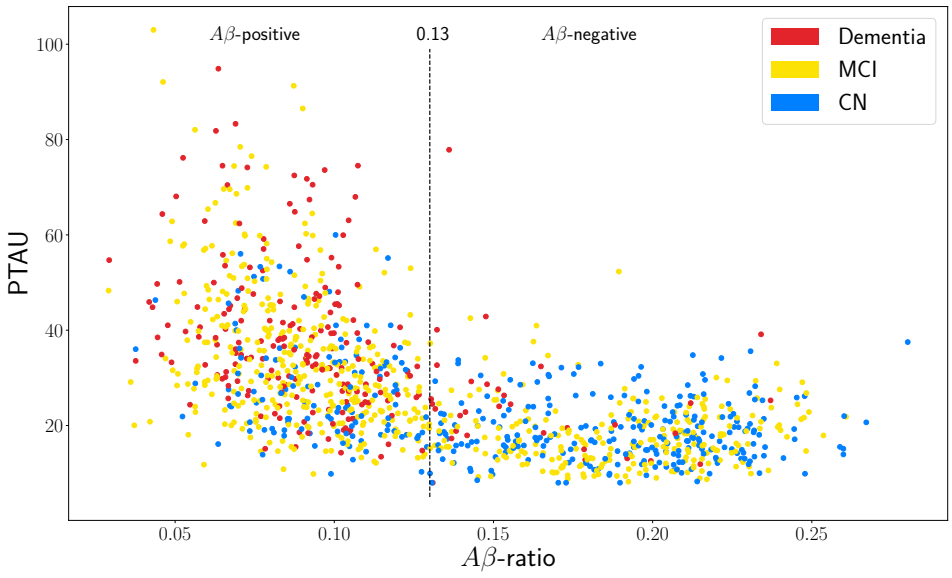

Figure 5.1: Scatter plot showing the $A\beta$-ratio versus PTAU of subjects at baseline. The different colors represent the different diagnoses. 0.13 is the ratio where patients are split into $A\beta$-positive and negative groups.

After splitting the cohort into groups of different diagnoses and $A\beta$ levels, violin plots were created for some of the most noticeable variables from literature. By

inspecting the violin plots in figure 5.2, one can see that for almost all of the plotted features, there is a noticeable difference between the $A\beta$-positive and negative groups when looking at all diagnoses together (right column of the figure). Looking at the different diagnoses groups (left column of the figure), several things can be observed. For example, in the age violin graphs, there are only small differences between the groups. Preferably, controls should have around the same distribution in age and sex and overall this seems fairly balanced. In the graphs with MMSE and ADAS13, it is evident that there is a difference between diagnoses as is expected, but also that in total the $A\beta$-positive group gets worse scores than the negative one (low MMSE scores and high ADAS13 scores are bad). In the $A\beta_{42}$ graph, a considerable separation is seen as would be expected but there is still some overlap between the two groups. The FDG is actually quite similar between the $A\beta$ groups although, in total, the positive group's scores are slightly lower. Furthermore, the average FDG is lower for those with dementia than CN and MCI. The APOE4 violins show that individuals with the gene are more likely to have a low $A\beta$ ratio and that the gene is more frequent in those with dementia than other diagnoses. An interesting observation is that almost no $A\beta$-negative subjects have two APOE4 genes. Comparing the AV45 and $A\beta_{42}$ graphs, the AV45 shows a slightly better separation between the positive and negative groups

For the CN and MCI groups, the cognitive progression over time was investigated. For each visit code, the groups' average MMSE cognitive test scores were calculated and plotted. The resulting graphs can be seen in figure 5.3. By looking at the graphs, a clear difference can be seen between the $A\beta$-groups for the CN subjects but especially the MCI subjects. The average score for the $A\beta$-positive MCI group four years after baseline was 23.79 while it started at 27.35 resulting in a drop in the average score of 3.56. On the other hand, the average score of the MCI $A\beta$-negative group started at 28.27 while after four years the average was 28.20 resulting in a drop in the average score of 0.07. The CN $A\beta$-positive and negative groups' changes in the average score were a drop of 0.8 and an increase of 0.08 respectively. Even though there are considerable differences in the average MMSE score deterioration between the $A\beta$ groups, it must be mentioned that subjects drop out of the study for various reasons. Therefore, the number of subjects is not the same at all points in the graph for each of the groups, i.e., there are fewer people still involved in the study after four years than at baseline. As an example of that, the number of people in the CN $A\beta$-positive and negative groups dropped from 119 and 248 at baseline to 69 and 142 after four years respectively and for the MCI $A\beta$-positive and negative groups they started at 370 and 248 people and after four years they were 167 and 149 respectively. Furthermore, no one was moved between groups over time in the figure even though they may have changed from $A\beta$-negative to positive or changed diagnosis during that time. This was because we were interested in the expected changes, only having information from the baseline. All in all, the amyloid negative groups seem to be more stable than the positive groups.

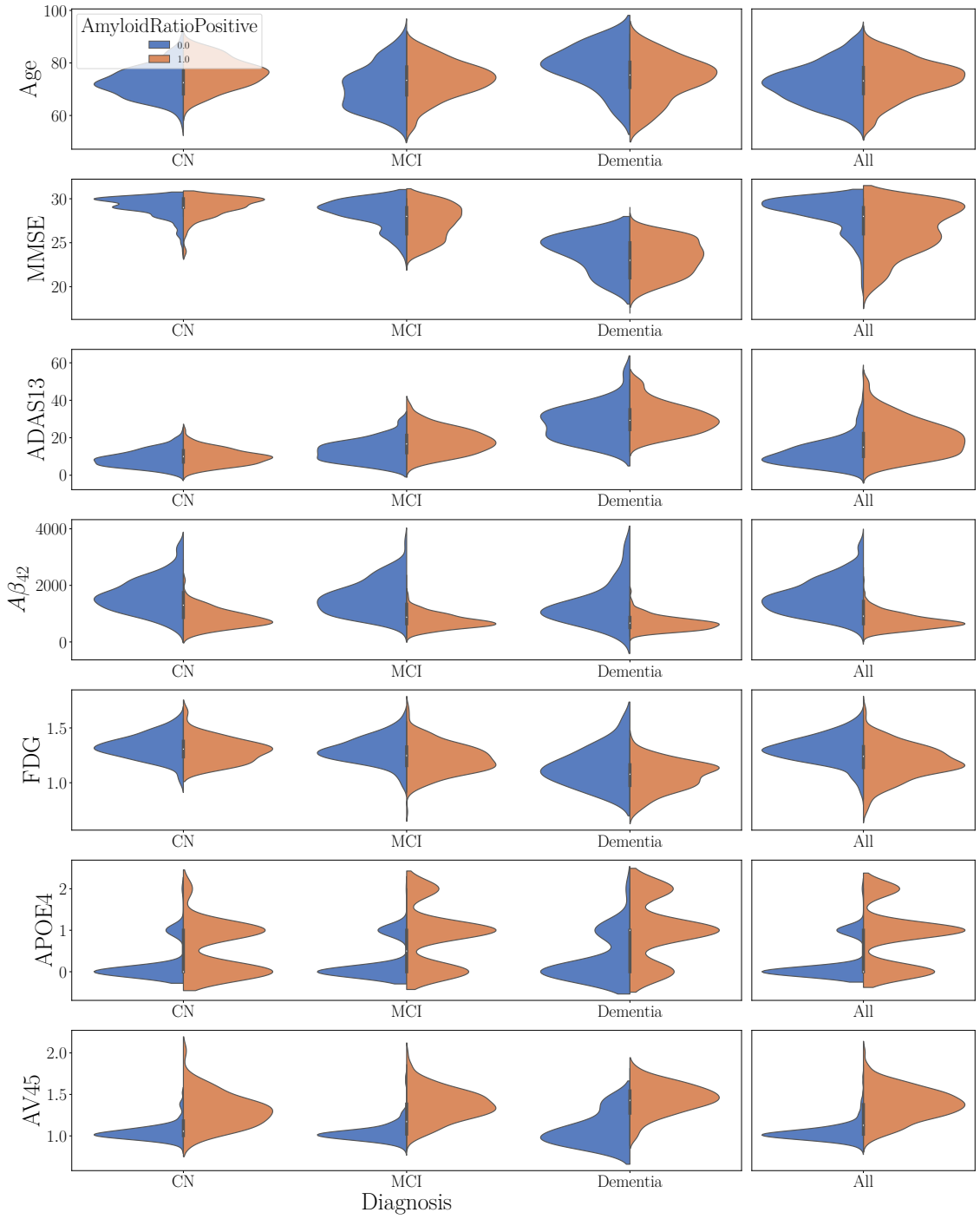

Figure 5.2: Violin plots of some of the variables frequently mentioned in literature on Alzheimer's. The left figure in each line shows the feature split by the three diagnoses but the right one shows all diagnoses together. Each plot is also split into $A\beta$-positive (orange) and negative (blue) parts. MMSE and ADAS13 are cognitive tests, FDG is a measure of the cerebral metabolic rate of glucose in the brain, $A\beta_{42}$ is a measure of $A\beta_{42}$ in CSF. APOE4 is a gene which is a risk factor for AD which people can have zero, one or two copies of. AV45 is a measure of the florbetapir mean of the whole cerebellum.

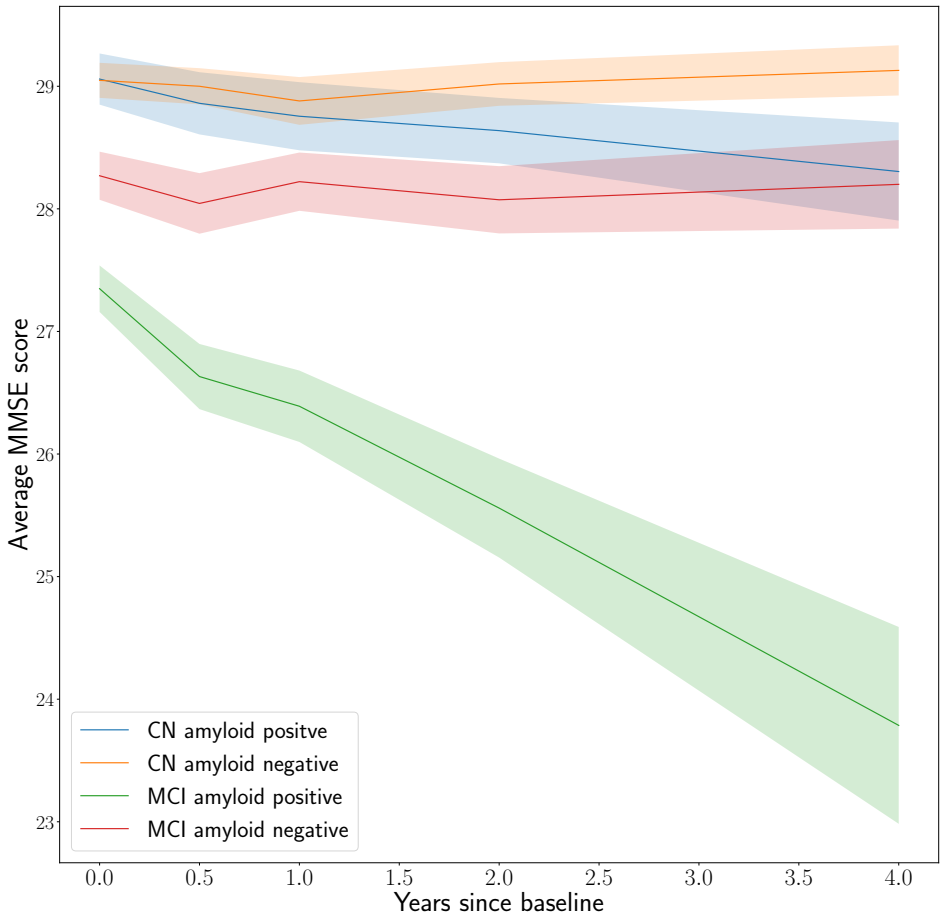

Figure 5.3: A graph showing how the MMSE score develops for CN and MCI subjects split into $A\beta$-positive and negative groups. The shaded areas represent the 95% confidence intervals for the mean values. It must be noted that the number of subjects drops down in every group as time passes.

5.2 Model results

In this part, the results from the feature selection and model performances are displayed. The results for each model type trained with different feature selections can be seen in table 5.1. For all predictions, an elastic net model with forward feature selection gives the best score. Generally, the elastic net models perform better than their random forest counterparts with the same feature selection method. Using all available features to train models without doing any prior selection results in a worse performance than if a good subset of features is selected. Using only a single feature also provides a much lower prediction score than if a good combination of features is used.

Table 5.1: Results from the models created using different feature selection methods. The regression models use the R^2 score while the classification models use the weighted F_1 score for evaluation. The values are average scores of the models trained using five different train/test splits mentioned in chapter 4. EN, RF and FFS stand for elastic net, random forest and forward feature selection respectively. The feature selection methods were FFS; using all features; using features selected by at least two different models when using FFS, which can be seen in figure 5.5; and single feature predictors. Out of the single predictors, only two are shown in the table, ADAS13 and FDG. Highlighted cells have a further visualisation of the model performance and feature importance later in the chapter.

	ADAS13 2 years, R^2	ADAS13 4 years, R^2	MMSE 2 years, R^2	MMSE 4 years, R^2	Classification weighted F_1
EN FFS (std)	0.514 (0.051)	0.428 (0.053)	0.580 (0.073)	0.577 (0.063)	0.817 (0.039)
RF FFS (std)	0.292 (0.050)	0.187 (0.230)	0.390 (0.056)	0.236 (0.126)	0.814 (0.037)
EN all features (std)	0.267 (0.104)	0.187 (0.090)	0.346 (0.082)	-0.523 (1.461)	0.721 (0.077)
RF all features (std)	0.249 (0.063)	0.148 (0.175)	0.324 (0.090)	0.168 (0.066)	0.702 (0.053)
EN chosen features (std)	0.378 (0.070)	0.252 (0.089)	0.361 (0.086)	0.292 (0.104)	0.754 (0.059)
RF chosen features (std)	0.315 (0.048)	0.224 (0.149)	0.353 (0.043)	0.253 (0.074)	0.757 (0.050)
EN ADAS13 (std)	0.121 (0.026)	0.035 (0.055)	0.274 (0.084)	0.211 (0.072)	0.695 (0.031)
RF ADAS13 (std)	0.083 (0.037)	-0.151 (0.188)	0.266 (0.056)	0.109 (0.145)	0.722 (0.017)
EN FDG (std)	0.110 (0.072)	0.127 (0.098)	0.110 (0.029)	0.107 (0.146)	0.627 (0.048)
RF FDG (std)	0.103 (0.085)	-0.090 (0.343)	0.117 (0.027)	0.067 (0.201)	0.572 (0.044)

Figure 5.4 shows how the score progresses during forward selection for the classification models. We can see that only a few features are needed to reach the best classification score. However, during forward selection for the elastic net regression models, they all reached the maximum of 25 additions without meeting the early stopping criterion but the last iterations added only a small increase. The features selected by forward selection vary between the different models. All features selected in two or more forward selections are shown in a heat map in figure 5.5 where the normalised importance of features is displayed. The first 7 features are cognitive tests and it can be seen that they are considered important predictors by the models. TAU and FDG also show up in many models and are given high importance. Other biomarkers also show up in a few models but are given lower importance in general. Indicators were not included in figure 5.5 even though they were used in training since they were generally assigned low importance by the models. The features selected by two or more models, displayed in figure 5.5, were used as a specific selection and all models were trained using these chosen features. Table 5.1 shows that this selection performs worse than the best models using forward selection but, in general, better than using all features.

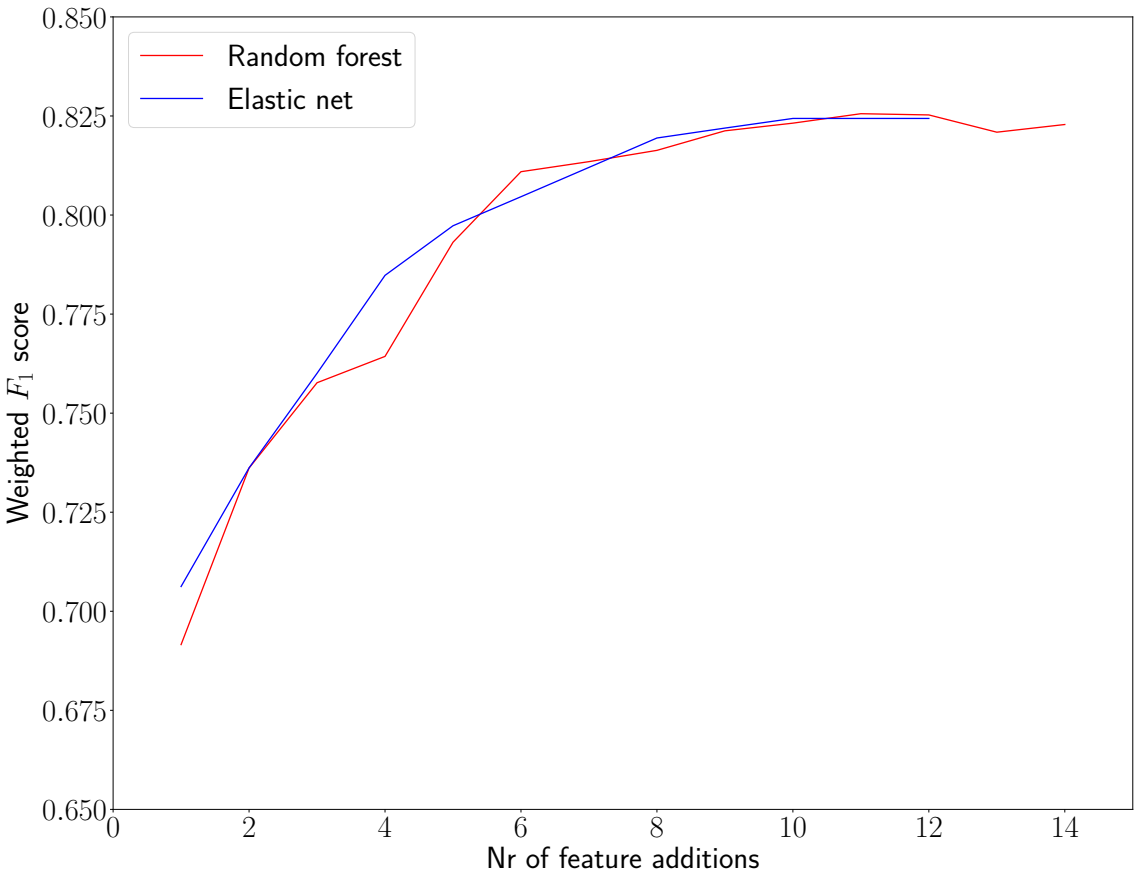

Figure 5.4: The F_1 scores obtained during forward selections after each addition for the classification models. The figure shows both the random forest and elastic net scores with the highest obtained at 12 and 10 additions respectively. Parameter search was performed after this selection.

.

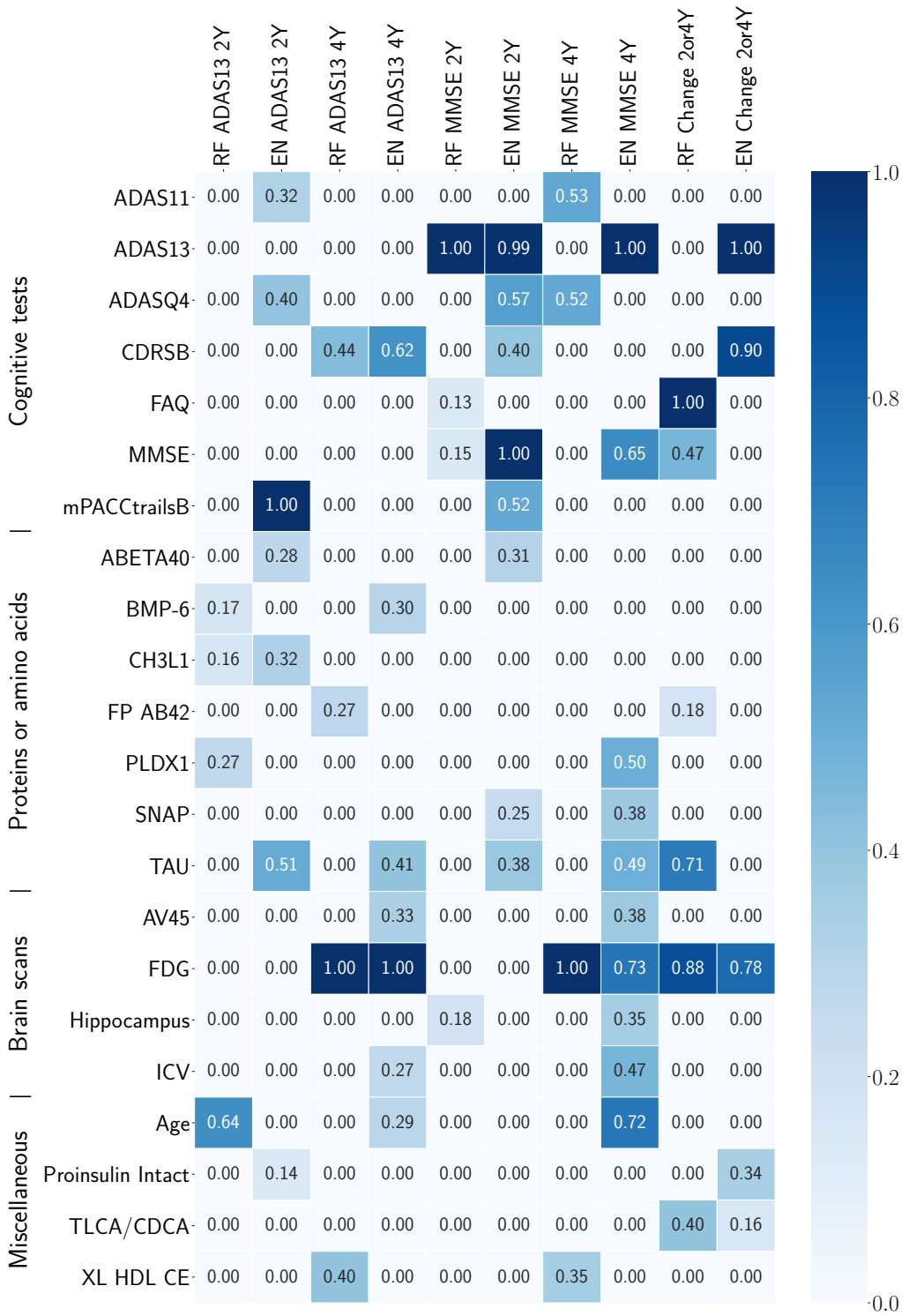

Figure 5.5: A heat map showing normalised feature importance of features which were forward selected by at least two models. For linear based models the importance is the absolute value of the linear coefficients. The y-axis has abbreviated feature names and the x-axis contains different model types. RF, EN and Y in the column names stand for random forest, elastic net and years respectively. The first seven features are cognitive tests, next seven are proteins or amino acids, next five are measures from brain scans and the rest do not fit in any of these groups. Feature explanations can be found in table B.1 in appendix B.

.

5.2.1 Regression: Change in ADAS13 score

The best cross-validated R^2 score for predicting change in ADAS13 score after two years was 0.514 with a standard deviation of 0.051. It was achieved using an elastic net regressor and 50 features selected. The features were chosen by forward selection reaching the maximum of 25 additions. The predicted values vs. true values can be seen in figure 5.6 as well as the linear coefficients. The figure shows one of the five models created and averaged over the 5-fold cross-validation. This is also the case in all figures displayed for other estimators. The highest contributing feature is mPACCtrailsB which is a cognitive test. The second-largest contribution is from TAU in CSF and the indicator for the availability of the TAU measurements is ranked third. The best random forest model received an average R^2 score of 0.315 with a standard deviation of 0.048. This was achieved using the 22 features and their missing indicators that were selected by two or more estimator types when using forward feature selection. Those features can be seen in figure 5.5 as mentioned earlier. The best single feature to predict the ADAS13 score after two years was mPACCtrailsB with a score of 0.236 and a standard deviation of 0.064 using elastic net. The elastic net using forward feature selection performed considerably better than all other model types using different feature selections.

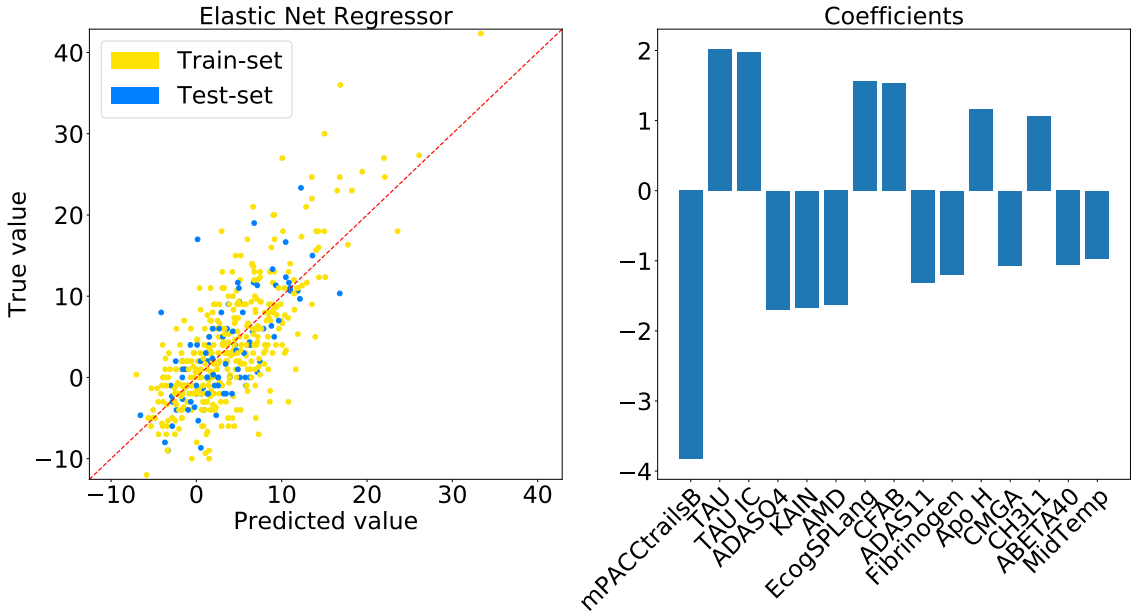

Figure 5.6: An elastic net model for predicting the change in ADAS13 score two years from baseline. The model uses forward selected features. The left side shows the predicted values versus the true values while on the right the linear coefficients of the model are displayed. The mPACCtrailsB is a cognitive test score, TAU is a protein measured in CSF and TAU IC is its missingness indicator. ADASQ4 is a cognitive test and KAIN is kallistatin protein measured in CSF. Explanations for other features can be found in table B.1 in appendix B.

The best cross-validated R^2 score for predicting change in ADAS13 after four years was 0.428 with a standard deviation of 0.053. As for the two years model, this was achieved by using elastic net using forward feature selection. 25 features were

selected along with missing indicators and one-hot encoded categories for a total of 53 features. The best random forest model achieved an average R^2 score of 0.224 with a standard deviation of 0.149. It used the features included by at least two models when using forward feature selection, outperforming its own forward selection. The most important feature for both model types was FDG which was also the best as a single predictor with a score of 0.127 and standard deviation of 0.098 using elastic net. An elastic net model trained using all features is shown in figure 5.7 as well as its linear coefficients. It shows one of the five models that give an average R^2 score of 0.187. It can be seen that it has trouble predicting the few individuals whose ADAS13 score has increased by over 20 points.

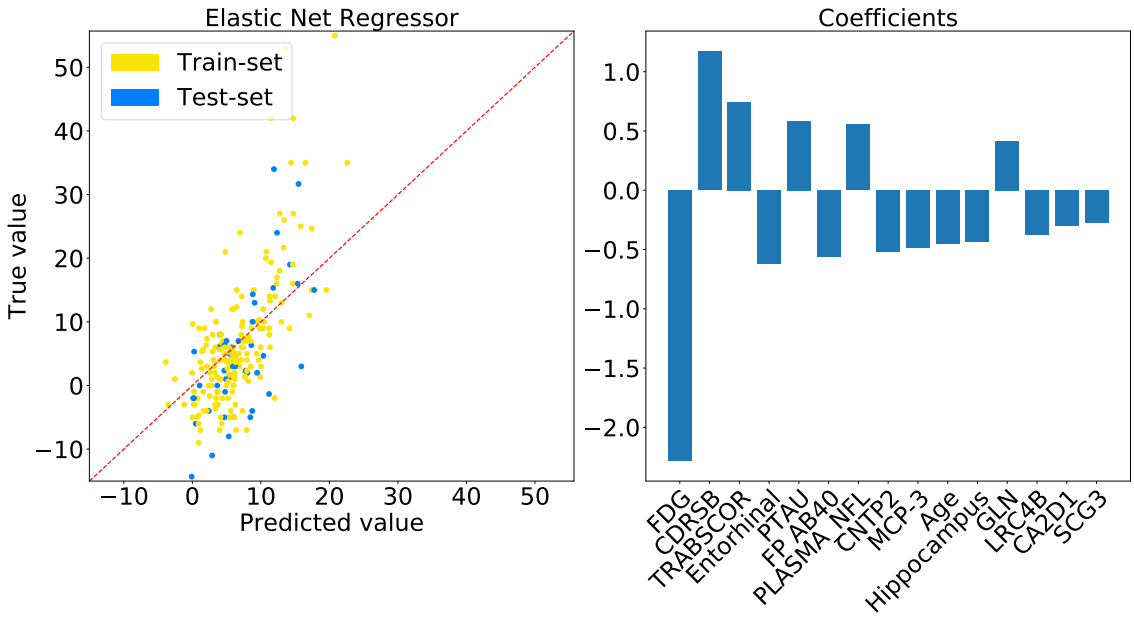

Figure 5.7: An elastic net model predicting the change in ADAS13 score after four years. The model is trained using all features. On the left: predicted values versus the true values. To the right: the linear coefficients of the model. FDG is a measure of the cerebral metabolic rate of glucose in the brain. CDRSB and TRABSCOR are cognitive tests. Entorhinal is a measurement of the size of a part of the brain. PTAU is a protein measured in CSF. Explanations of other variables can be found in table B.1 in appendix B.

5.2.2 Regression: Change in MMSE score

The best cross-validated R^2 score for predicting change in MMSE two years after baseline was 0.580 with a standard deviation of 0.073. It was achieved using an elastic net regressor and 53 features. The features were selected by forward selection. The best random forest model achieved an R^2 score of 0.390 with a standard deviation of 0.056 also using forward selection but only selecting 20 features. The random forest model using all 22 features picked by at least two models with forward feature selection resulted in a slightly lower score of 0.353 with a standard deviation of 0.043. The predicted values versus true values as well as feature importance of one such model can be seen in figure 5.8. Comparing the difference between the

true and predicted values of the train and test set, it is evident that the model has been overfitted slightly as the yellow points are closer to the diagonal. As for the prediction of ADAS13 change after four years, there seem to be a few outliers whose values are harder to predict. The best single predictor for the MMSE score after two years was ADAS13. When using only ADAS13 for prediction, the scores were similar for both elastic net and random forest averaging in 0.274 and 0.266 with standard deviations of 0.084 and 0.056 respectively. Cognitive tests gave the highest importance in all models predicting MMSE score change after two years where ADAS13 was the most dominant one.

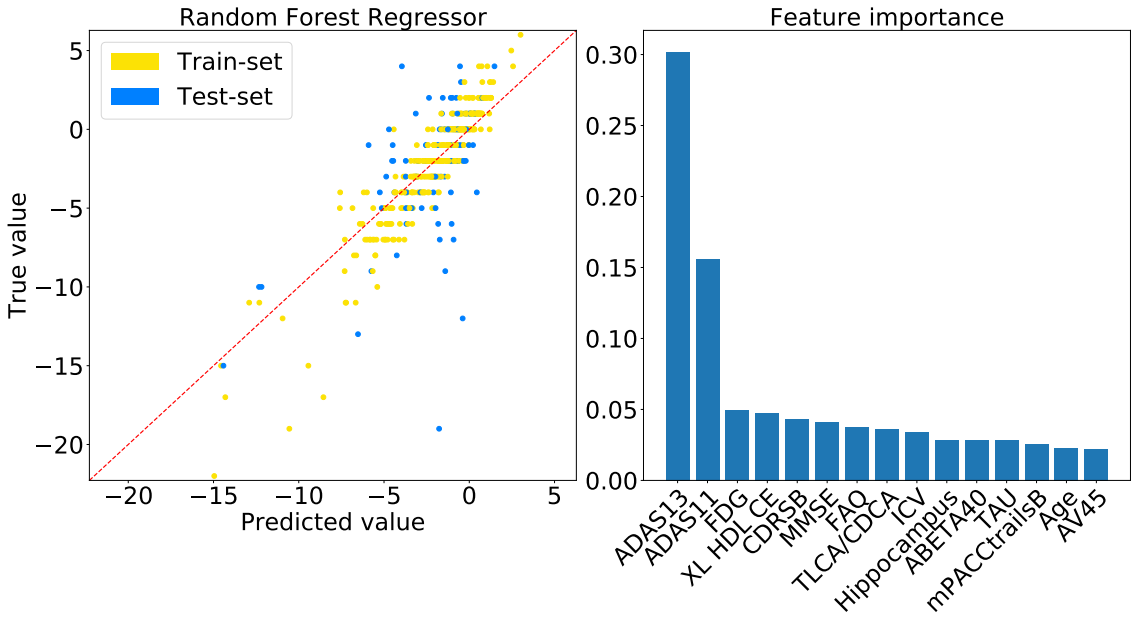

Figure 5.8: A random forest model for predicting the change in MMSE score two years from baseline. The model is trained using the features chosen by at least two models in forward selection. The left side shows the predicted values versus the true values while the right one shows the estimated feature importance of the model. ADAS13, ADAS11 and CDRSB are cognitive test scores. FDG is a measure of the cerebral metabolic rate of glucose in the brain. XL HDL CE is cholesterol esters in very large high-density lipoproteins. Explanations of other variables can be found in table B.1 in appendix B.

The best cross-validated R^2 score for predicting change in MMSE after four years was 0.577 with a standard deviation of 0.063. It was achieved using elastic net with forward feature selection using 50 features in total. In figure 5.9, one of the models contributing to the best average is shown. The predicted versus true values are shown as well as the linear coefficients of the top 15 features. The few individuals having an MMSE score decrease of over 10 are all predicted to have decreased less. The most important features, i.e., that have the largest absolute linear coefficients are ADAS13 and FDG. The best random forest received an R^2 score of 0.253 with a standard deviation of 0.074. It was achieved using the 22 features chosen by at least two models with forward feature selection along with their missingness indicators. One of the averaged models using elastic net with all features performs so poorly

that the average score becomes negative with a standard deviation of 1.461. The best single predictor for predicting the MMSE score after four years was ADAS13 with a score of 0.211 and standard deviation of 0.072 using elastic net. The elastic net models using forward feature selection performed considerably better than all others in predicting the progression of the MMSE score as can be seen in table 5.1.

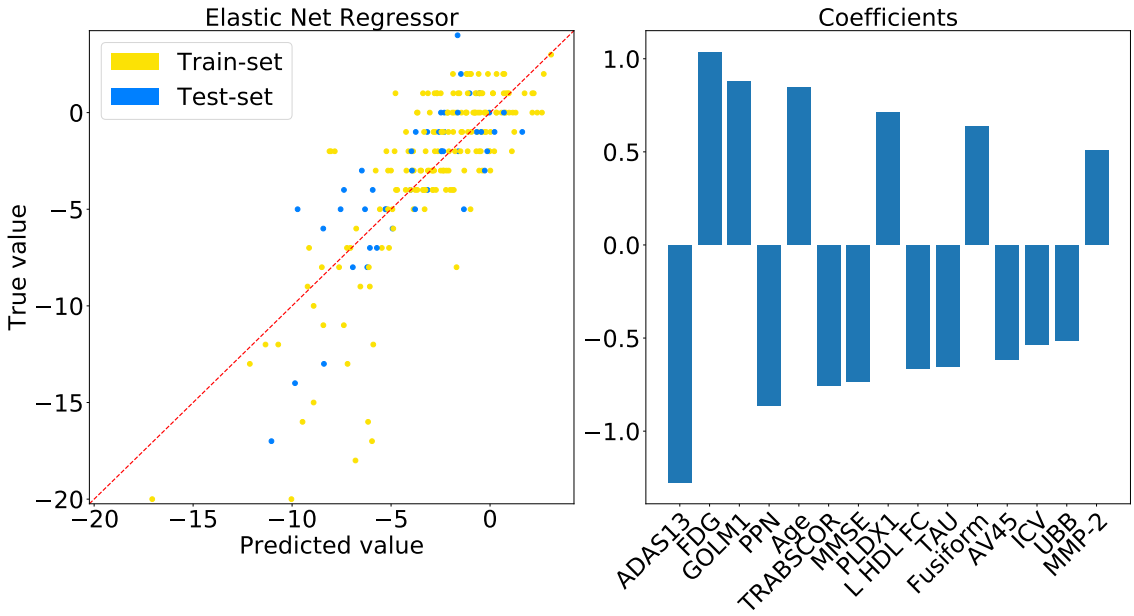

Figure 5.9: An elastic net model for the change in MMSE score after four years. The model is trained using forward selected features. On the left: predicted values versus true values. To the right: the feature importance of the model. ADAS13 is a cognitive test score. FDG is a measure of the cerebral metabolic rate of glucose in the brain. GOLM1 is the golgi membrane protein 1 measured in CSF. PPN is a measurement of papilin protein in CSF. Age is the age of a subject at baseline. Explanations of other variables can be found in table B.1 in appendix B.

5.2.3 Classification: Worse diagnosis in two or four years?

The best model created for the binary classification of whether or not a person's diagnosis changes after two or four years was made with logistic regression with elastic net penalty, resulting in a cross-validated weighted F_1 score of 0.817 with a standard deviation of 0.039. It used 20 features, 10 of which were indicators. The features were selected using forward feature selection. Results from one of the five models made using different test/train splits is shown in figure 5.10. The figure shows the confusion matrices for both the test and the train sets as well as the linear coefficients of the model. The model performs similarly on the training and test set and the features it considers most important are ADAS13, CDRSB and FDG. The accuracy of the model was 0.816 with a standard deviation of 0.040.

The best random forest model performed almost equally good with a cross-validated F_1 weighted score of 0.814 with a standard deviation of 0.037. It uses 23 features chosen by forward feature selection where 12 of them are indicators or one-hot

encoded values. The models using forward feature selection were both better than other models created.

When using all features, the resulting F_1 weighted scores were 0.721 and 0.702 with standard deviations of 0.077 and 0.053 for logistic regression and random forest respectively. Results from one of the five random forest classifiers using all features are shown in figure 5.11. By inspecting the confusion matrices for the test and train set, the model seems to be overfitted to the training set. The best single predictor was ADAS13 using random forest. The average F_1 weighted score for it was 0.722 with a standard deviation of 0.017. When using all features selected by two or more models using forward feature selection, the scores were 0.754 for logistic regression and 0.757 for random forest with standard deviations of 0.059 and 0.050 respectively which is lower than the forward selection but higher than using all of them or only a single feature.

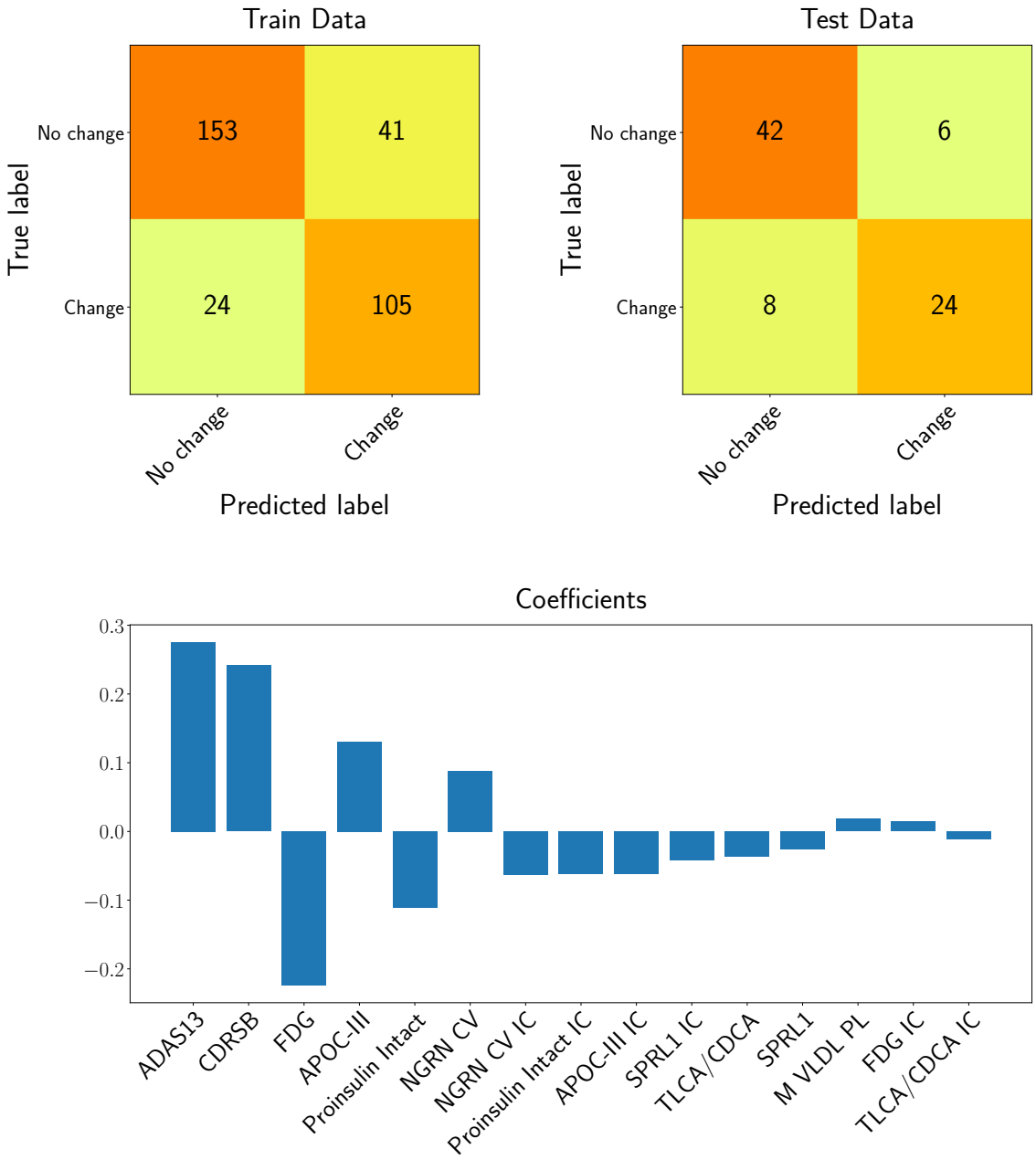

Figure 5.10: Confusion matrices and linear coefficients for the best classification model using logistic regression. The figure shows the results of a model using one of the five train/test splits. ADAS13 and CDRSB are cognitive tests. FDG is a measure of the cerebral metabolic rate of glucose in the brain. APOC-III is a protein encoded by the APOC3 gene and proinsulin intact predicts progression of insulin resistance. Information on other features shown in the figure can be found in table B.1 in appendix B.

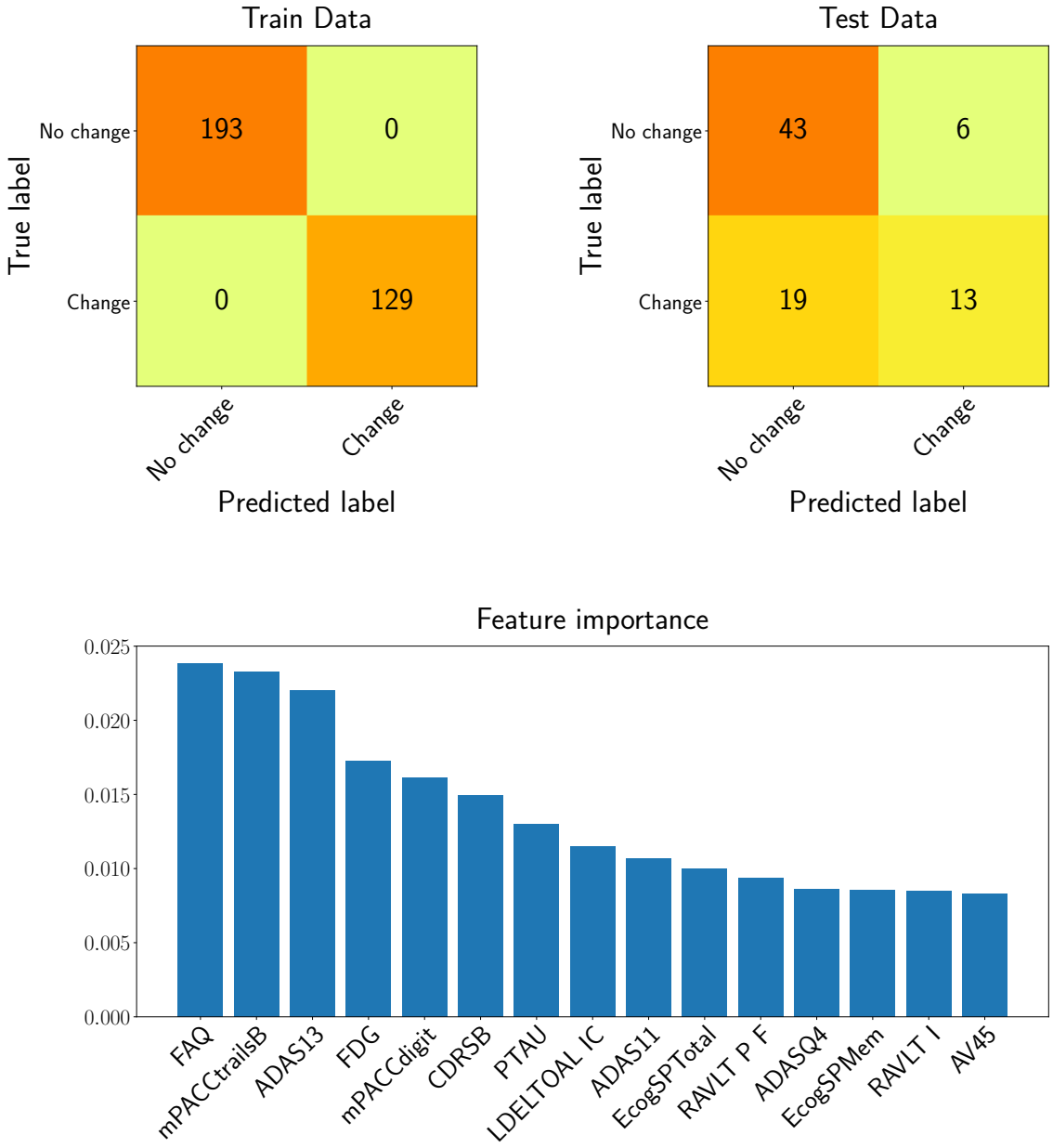

Figure 5.11: Confusion matrices and feature importance for a random forest using all features. The figure shows the results of a model using one of the five train/test splits. The confusion matrices show that the model is overfitted to the training data and does not generalise well. FAQ, mPACCtrailsB, ADAS13 and mPACCdigit are cognitive test scores. FDG is a measure of the cerebral metabolic rate of glucose in the brain. Information on other features in the lower plot can be found in table B.1 in appendix B.

6

Discussion

6.1 Data exploration

Initial analyses which include figures 4.1, 4.2 and 5.1 show that there seem to be two groups of subjects based on differences in the $A\beta$-ratio. The exact split between the groups is, nonetheless, not obvious and in figure 4.2 they seem to overlap but have a fairly good split at 0.14 - 0.15. However, figures 4.1 and 5.1 show that based on diagnosis and cognitive decline perhaps it should be slightly lower resulting in the choice of 0.13. Possibly, the ratio split should be set slightly higher but this is difficult to decide precisely and is done using visual inspection of the figures mentioned before. Dementia at baseline is very rare in the $A\beta$-negative group and may be caused by other diseases than AD.

The violin plots in figure 5.2 show that there are differences between the two $A\beta$ groups and also between subjects with different diagnoses. The visible separation between $A\beta$-positive and negative groups in the $A\beta_{42}$ graph is expected. However, the $A\beta$-negative group has a much larger spread which indicates that some subjects have low $A\beta_{42}$ but also low $A\beta_{40}$. This might mean that they do not have AD pathology but are just low producers of $A\beta$ in general. As mentioned in the results, the AV45 shows a slightly better separation between the positive and negative groups than the $A\beta_{42}$ graph, indicating that perhaps the ratio agrees better with the PET imaging of $A\beta$ deposits in the brain than using just $A\beta_{42}$ in CSF.

It is worth mentioning that even though $A\beta$-positivity by PET is usually highly correlated with low CSF-$A\beta_{42}$ levels in people with AD, some people, usually CN, with similarly low CSF-$A\beta$ levels are $A\beta$-negative by PET [9]. Even though the AV45, which measures $A\beta$ in PET, is included in the data, it had fewer measurements compared to CSF-$A\beta$. Therefore, $A\beta$-positivity is only derived from CSF measures in this project. Furthermore, these individuals are often considered as outliers by researchers in the field, so it should be reasonable to use the CSF measure in this project.

When looking at figure 5.3, a visible difference in the drop of the average MMSE score between the $A\beta$-positive and negative groups can be observed. However, as mentioned in section 5.1, the fact that subjects drop out over time must not be overlooked. The reasons for subjects dropping out of the study are unknown. This

might produce some bias as the people dropping out are possibly doing considerably worse and some of the graphs could, therefore, end at a slightly higher average than if the data were complete but the opposite might also apply. However, the dropout rate of people is around 40% in all groups except the MCI $A\beta$-positive group where it is 55%. It is therefore likely that the differences between the groups in the figure are existent. Since the drop in MMSE score is the highest for the $A\beta$-positive MCI group, it also seems a likely assumption that those who deteriorate more have a higher risk of dropping out of the study. Possibly, some have developed AD dementia and the reduced life expectancy of people with AD [16] could be the reason for some of them dropping out.

6.2 Model performance

In all models, at least one cognitive test showed up as one of the most important features for prediction. In many cases, more cognitive tests were included as important features. However, adding more and more cognitive tests does not seem to improve the results much. This is probably because the test scores are highly correlated and do not add much value when other cognitive tests are already in use. The correlation between cognitive tests selected by two or more models in forward feature selection is shown in a heat map in figure 6.1. Because of this, it might be of some value to create a combined cognitive score to use as a single feature. It seems apparent that cognitive test results indicate how the individual will progress and that those who already score worse are more likely to be worse off in the future.

No feature shows up in all 10 forward feature selections as can be seen in figure 5.5. FDG shows up most frequently, being part of six selections, TAU is second most frequent and is selected five times while ADAS13, MMSE and CDRSB are all selected four times. This indicates that there is not a single feature that accurately predicts progression on its own although it looks like FDG is the biggest indicator of decline after four years. FDG on its own however performs significantly worse than a good combination of features. TAU also shows up in all forward feature selections within elastic net regression models but no random forest ones, implying that it has a small linear relationship which does not provide clear enough separations based on gini impurity. Both FDG and TAU proteins have been researched extensively in the context of AD and are both considered indicators of neurodegeneration in the AT(N) model [11]. We did not find new features that seemed to shield individuals from deteriorating.

After viewing the data, including subjects of any $A\beta$-ratio, it would be expected that features based on the magnitude of $A\beta$ might have higher importance within the models. However, after choosing only those who are $A\beta$-positive, $A\beta$ based features are given rather low importance. Thus, a binary categorisation of the $A\beta$-ratio seems like a reasonable approach.

Many proteins, lipids, brain scans and other biomarkers only show up in two or fewer selections and are given normalised importance of under 0.5. Such features sometimes have a rather low proportion of subjects with measurements. For example,

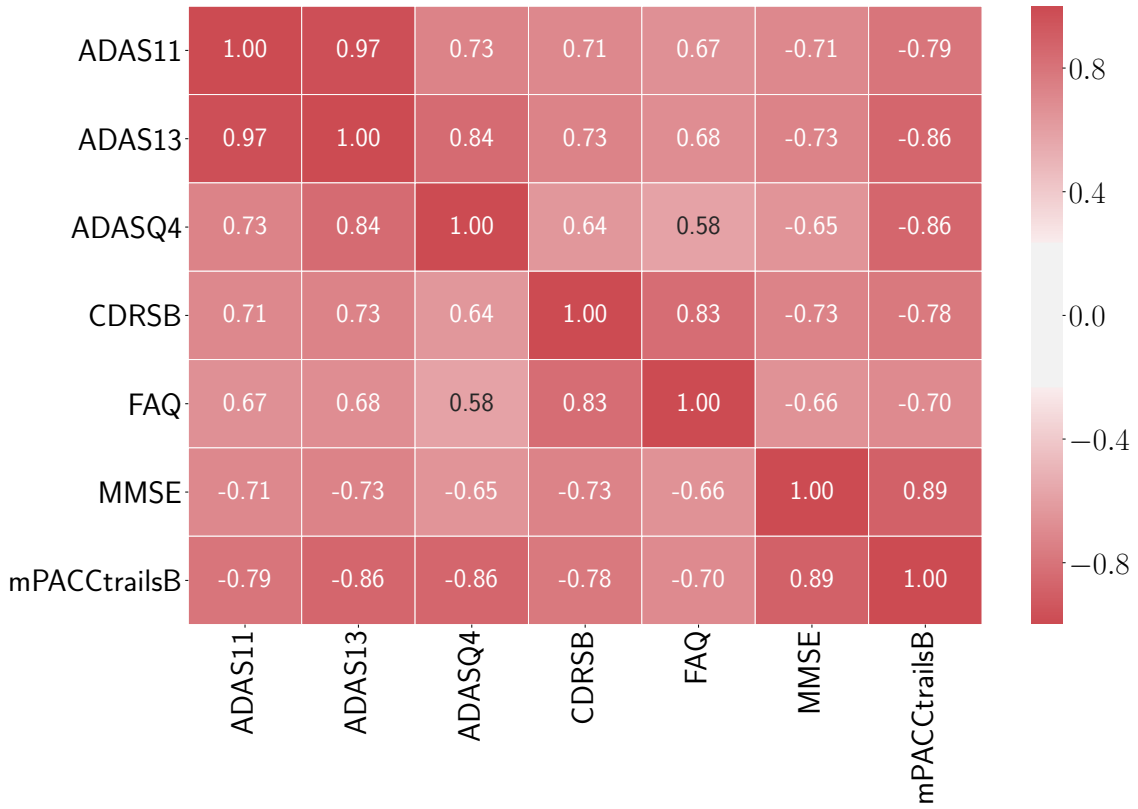

Figure 6.1: A heat map showing the correlation between the cognitive tests that were selected by two or more models in forward selection. Darker red colours indicate more correlation, either negative or positive

BMP-6, CH3L1, PLDX1, SNAP and Proinsulin Intact all have under 570 measurements at baseline out of the possible 2253. To get a better estimate of their roles and predictive capabilities it would be beneficial to fill in more of the gaps in measurements. They might also benefit from using a different imputation strategy for missing values such as multiple imputation but such a strategy could also hurt the performance or create false patterns.

6.2.1 Regression

The elastic net model using forward feature selection performed best for all regression tasks. Thus, a linear relationship between the features and the outcome appears to be a reasonable assumption and a more flexible model which can learn more complex relationships like the random forest regressor may be prone to overfitting to patterns that are not there. When viewing figure 5.8 of a random forest model, we can see that the training data are fitted quite well but the test data perform significantly worse indicating that the model is indeed overfitted. It is interesting to see that, in general, the random forest regression models perform worse when using their own forward feature selection than when choosing the features that are selected by at least two models. This shows that the forward selection does not find the best feature set and the selection could perhaps be improved, such as by allowing

more additions without improvement before early stopping or allowing deletions of previously selected features if it would improve the score. The elastic net regressors all achieve their best results after doing the maximum of 25 additions which means that perhaps a higher score could be reached. It would, therefore, be better to allow the forward selection to be carried out until the score starts to decline but it was computationally heavy to carry out so many selections and as the increase was small after 20 additions the maximum was not increased further.

Our models seem to be slightly better at predicting MMSE score progression than that of ADAS13. For all regression targets, there are a few outliers whose change in score is much larger than others. These outliers can be difficult to predict and we can see in figures 5.7, 5.8 and 5.9 that the models struggle to properly predict their scores and these outliers may potentially decrease the quality of predictions of other data points. It may, therefore, be possible to increase the quality of other predictions if the effects of these outliers are mitigated. It also seems easier to predict ADAS13 score after two years than after four years but the best score for MMSE change is similar when predicting two or four years. It is quite interesting that the performance is similar when predicting further into the future even though there should be more uncertainty. This might indicate that if the subjects are set on a certain path of progression they are likely to follow it quite closely. The number of subjects also decreases substantially after four years compared to two years which is also expected to negatively impact the generalisation capabilities of the models.

6.2.2 Classification

The best classification model created was the logistic regression model with elastic net penalty using forward feature selection and hyperparameter search. However, the random forest model performed almost equally good with a difference of only 0.003 in the weighted F_1 score and with 0.002 less standard deviation. When all classification models are considered, the forward feature selection models perform considerably better than models using other feature selection methods. This is true for both random forest and logistic regression for classification while for regression, only the elastic net performs best with forward feature selection. No single feature has as good predictive power as the set of features chosen by forward feature selection. This suggests that several features play a part in how a person's diagnosis develops and no single feature is the only cause, at least of the ones included in the data set. Furthermore, models using all available features performed considerably worse than the best models. This can be because using all features may lead to overfitting of the models to the training set like evidently happened for the random forest model whose results are shown in figure 5.11. In this case, the model has been fit perfectly to the training set but contrarily does not perform well on the test set. That shows that it has not generalised well.

Unlike for the regression models, neither of the model types, elastic net nor random forest, seems to be better at classifying using all feature selection methods. Random forest performs better with the chosen features and when only using ADAS13 but in the other three cases shown in table 5.1, elastic net shows better results. In no case

does one of the models significantly outperform the other type. All in all the forward feature selection models perform best when deciding whether or not a person will deteriorate so that they get a worse diagnosis within four years.

6.3 Future work

The possible directions to take in this project are abundant. A few feasible next steps are discussed below that would be interesting to investigate further with extended time and resources.

6.3.1 Genetic data

The data from ADNI include the full genome of the patients. It would be interesting to include these in the models and see if some unknown patterns in the genome can be associated with AD progression. However, these data are huge, e.g., SNPs data are around 52 GB and the whole genome for all patients is around 150 TB. Furthermore, expert knowledge within the field of genetics would be required to extract the relevant information from the data.

6.3.2 Temporal spatial information for FDG

FDG is a radiopharmaceutical used in PET scans. The uptake of FDG in the brain provides estimates of the cerebral metabolic rate of glucose (CMRglc), which is a direct indicator of synaptic functioning and density [30] and can be measured in different parts of the brain. The FDG feature in the ADNI data set is the average measurement of FDG-PET in the angular gyrus, temporal lobe, and posterior cingulate cortex. However, it might be of value to include single region FDG-PET measures since different types of dementia seem to show hypometabolism in different regions of the brain [25].

6.3.3 Multi-class classification

It would be interesting to create a multi-class classifier instead of a regressor where the classes would be some interval of change in cognitive score, for example, no change or positive change; 1-5 points decline; 6-10 points decline; and 10 points or more decline for the MMSE and something comparable for ADAS13. This could be an easier task for a prediction model since the regression is predicting an accurate score while the classifier would have more space for small errors. Furthermore, this method would help with the problem introduced by outliers in the data where the score of a few subjects changes drastically. It would still give the user much value since he or she would be interested in whether or not the change will be big. It is probably more important to know whether a big decline is predicted rather than to know the exact score change. One or two point difference might not matter. The cognitive ability, in general, is what is of interest but not the exact score value.

6.3.4 Inexpensive or easily obtained features only

Some methods for obtaining biomarkers are more expensive or invasive than others. For example, PET scans are more expensive than blood samples and CSF fluid extraction is more invasive since it requires a lumbar puncture. Creating a model only using features that are inexpensive to measure or are non-invasive for patients could be something to consider. It would be interesting to see whether it is possible to achieve similarly good estimators when excluding the more expensive measures like PET scan data and the more invasive ones like CSF data. If that was the case, a risk score for the cognitive decline could perhaps even be conducted in a routine doctor checkup or screening for the elderly.

6.3.5 Different imputation method

While the imputation method used here is computationally simple and makes it possible to use the non-complete data, it does not take information from other features into account that may be relevant. Imputing the values as the feature's mean may result in underestimation of the variability of the unseen data [8]. Instead, multiple imputation could be used. Using this method means generating m complete sets of the data, each with different but plausible imputed values for the missing observations, and analysing each data set separately. For this method to work well, features with missing observations should be predictable from available information which might not be the case for all features. Additionally, it is more computationally heavy than the simple imputation we use. However, it might be of some value since we then have a variance for these missing features. Possibly, this imputation method would improve our models' performances.

6.4 Limitations

The data are not complete as has been mentioned before. This is a limiting factor in this project. For many features, we have imputed the missing values to avoid the limitations and possible biases the missingness could create. Furthermore, by doing this we can include more lines of data even though they do not all include original values for all features. However, we do not impute the output variables for the model. The amount of data is limited by this since for each model, we only include individuals who have the output variables available. If these data were complete, the models might be more powerful and perform better since they would have more data to learn from.

Another limitation present in the data is the fact that the diagnoses are CN, MCI and dementia, i.e., dementia as a general term and AD dementia is not specified as one specific diagnosis. Many types of dementia exist that are not related to AD. However, one file in ADNI includes an indicator of whether or not the diagnosed dementia is caused by AD or not. When comparing the $A\beta$-positive patients diagnosed with dementia, the ones whose AD indicator is available (which is around half) are all marked as having AD. However, for the rest, we do not know for certain whether

they are diagnosed with AD dementia or some other kind of dementia. Because of this, our classifiers cannot be said to predict whether or not a person gets AD dementia. On the other hand, only $A\beta$-positive patients are considered, eliminating the $A\beta$-negative ones who most likely do not have AD dementia.

Furthermore, the project is limited by the data gathered. The data do not include much information on the subjects' lifestyles, drugs they use or accidents of some kind they may have been in. These factors might be the cause of some cognitive decline or even prevent the decline from happening. We do not know the effect of these factors and cannot capture any correlation between them and the cognitive abilities of patients since these data are not available in the data set. Some of these things may also be factors for why patients drop out of the study. Reasons for the dropouts are not known and if everyone would continue the study, the results might differ. Furthermore, the data in ADNI is from America only. Therefore, it is possible that this data set cannot be generalised as patient data for the whole world.

7

Conclusion

The $A\beta$-positive and negative patients are quite clearly distinguishable. The $A\beta$-positive patients seem more likely to get a worse diagnosis within four years than the $A\beta$-negative ones as is shown in figure 4.1. Furthermore, figure 5.3 supports this claim, since the average MMSE score for the $A\beta$-positive groups seems to decline more within four years from the baseline. The difference between the two groups is also clear when looking at some of the features at baseline as the violins in figure 5.2 show when plotting all diagnoses groups together.

The five different estimators show promising results. The best performance is achieved using elastic net with forward feature selection for all estimators. The best regressors have average R^2 scores between 0.428 and 0.580 while the best classifier has an F_1 weighted score of 0.817. We, therefore, conclude that it is possible, to some extent, to estimate the rate of progression of people with indicators of $A\beta$ plaques. With further development, we see the potential for these kinds of predictions to assist in clinical settings.

Different features are important for the estimators. The most common ones, chosen by at least two of the ten models are shown in figure 5.5. The ones that seem of most importance in general are FDG, TAU and a few cognitive tests including ADAS13 and MMSE. No single predictor performs as good as when using a good selection of several features. Thus, no feature in the data seems to be the single reason for some $A\beta$-positive patients deteriorating faster than others. The features that are found to be of high importance by the models have been researched extensively in the context of AD, i.e., they have previously been associated with Alzheimer's.

Bibliography

[1] Edoardo Amaldi, Viggo Kann, et al. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209(1):237–260, 1998.

[2] John Barnard and Xiao-Li Meng. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical methods in medical research*, 8(1):17–36, 1999.

[3] Kaj Blennow, Niklas Mattsson, Michael Schöll, Oskar Hansson, and Henrik Zetterberg. Amyloid biomarkers in Alzheimer's disease. *Trends in pharmacological sciences*, 36(5):297–309, 2015.

[4] Heiko Braak and Eva Braak. Neuropathological stageing of Alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259, 1991.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Hi Crystal, D Dickson, P Fuld, D Masur, R Scott, M Mehler, Joseph Masdeu, C Kawas, M Aronson, and L Wolfson. Clinico-pathologic studies in dementia: nondemented subjects with pathologically confirmed Alzheimer's disease. *Neurology*, 38(11):1682–1682, 1988.

[7] JP Dick, RJ Guiloff, A Stewart, J Blackstock, C Bielawska, EA Paul, and CD Marsden. Mini-mental state examination in neurological patients. *Journal of Neurology, Neurosurgery & Psychiatry*, 47(5):496–499, 1984.

[8] James D Dziura, Lori A Post, Qing Zhao, Zhixuan Fu, and Peter Peduzzi. Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale journal of biology and medicine*, 86(3):343, 2013.

[9] Anne M. Fagan. What does it mean to be 'amyloid-positive'? *Brain*, 138(3):514–516, 02 2015.

[10] Alzheimer's Disease International. About dementia. `https://www.alz.co.uk/about-dementia`. Accessed: 2020-02-04.

[11] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank

Jessen, Jason Karlawish, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

[12] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37, 2000.

[13] Kristel JM Janssen, A Rogier T Donders, Frank E Harrell Jr, Yvonne Vergouwe, Qingxia Chen, Diederick E Grobbee, and Karel GM Moons. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*, 63(7):721–727, 2010.

[14] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[15] Jacqueline K. Kueper, Mark Speechley, and Manuel Montero-Odasso. The Alzheimer's disease assessment scale–cognitive subscale (ADAS-Cog): modifications and responsiveness in pre-dementia populations. a narrative review. *Journal of Alzheimer's Disease*, 63(2):423–444, 2018.

[16] Eric B Larson, Marie-Florence Shadlen, Li Wang, Wayne C McCormick, James D Bowen, Linda Teri, and Walter A Kukull. Survival after initial diagnosis of Alzheimer disease. *Annals of internal medicine*, 140(7):501–509, 2004.

[17] Piotr Lewczuk, Hermann Esselmann, Markus Otto, Juan Manuel Maler, Andreas Wolfram Henkel, Maria Kerstin Henkel, Oliver Eikenberg, Christof Antz, Wolf-Rainer Krause, Udo Reulbach, et al. Neurochemical diagnosis of Alzheimer's dementia by CSF Aβ42, Aβ42/Aβ40 ratio and total tau. *Neurobiology of aging*, 25(3):273–281, 2004.

[18] Piotr Lewczuk, Anja Matzen, Kaj Blennow, Lucilla Parnetti, Jose Luis Molinuevo, Paolo Eusebi, Johannes Kornhuber, John C Morris, and Anne M Fagan. Cerebrospinal fluid Aβ 42/40 corresponds better than Aβ 42 to amyloid PET in Alzheimer's disease. *Journal of Alzheimer's Disease*, 55(2):813–822, 2017.

[19] Chia-Chen Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2):106, 2013.

[20] Xiaoli Liu, Peng Cao, Jinzhu Yang, and Dazhe Zhao. Linearized and kernelized sparse multitask learning for predicting cognitive outcomes in Alzheimer's disease. *Computational and mathematical methods in medicine*, 2018, 2018.

[21] Donald W Marquardt and Ronald D Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

[22] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M Stadlan. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department

of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, 34(7):939–939, 1984.

[23] Carlos O Mendivil, Chunyu Zheng, Jeremy Furtado, Julian Lel, and Frank M Sacks. Metabolism of VLDL and LDL containing apolipoprotein C-III and not other small apolipoproteins–R2. *Arteriosclerosis, thrombosis, and vascular biology*, 30(2):239, 2010.

[24] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Alzheimer's Disease Neuroimaging Initiative, et al. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*, 104:398–412, 2015.

[25] Lisa Mosconi, Wai H Tsui, Karl Herholz, Alberto Pupi, Alexander Drzezga, Giovanni Lucignani, Eric M Reiman, Vjera Holthoff, Elke Kalbe, Sandro Sorbi, et al. Multicenter standardized 18F-FDG PET diagnosis of mild cognitive impairment, Alzheimer's disease, and other dementias. *Journal of nuclear medicine*, 49(3):390–398, 2008.

[26] 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*, 15(4):869–877, 2005.

[27] National Institude of Aging. What are the signs of Alzheimer's disease? `https://www.nia.nih.gov/health/what-are-signs-alzheimers-disease`, May 2017. Accessed: 2020-04-27.

[28] World Health Organization. Dementia. `https://www.who.int/news-room/fact-sheets/detail/dementia`. Accessed: 2020-02-05.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[30] Anne B Rocher, Françoise Chapon, Xavier Blaizot, Jean-Claude Baron, and Chantal Chavoix. Resting-state brain glucose utilization as measured by PET is directly related to regional synaptophysin levels: a study in baboons. *Neuroimage*, 20(3):1894–1898, 2003.

[31] Shai Shalev-Shwartz and Shai Ben-David. *Feature Selection and Generation*, page 309–322. Cambridge University Press, 2014.

[32] Claudio Soto. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nature Reviews Neuroscience*, 4(1):49–60, 2003.

[33] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[34] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, et al. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

[35] Ian R White and Simon G Thompson. Adjusting for partially missing baseline measurements in randomized trials. *Statistics in medicine*, 24(7):993–1007, 2005.

[36] Fen Xia, Wensheng Zhang, Fuxin Li, and Yanwu Yang. Ranking with decision tree. *Knowledge and information systems*, 17(3):381–395, 2008.

[37] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.

[38] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

A

Appendix A

Parameter	Regression	Classification
No. of estimators	50, 100, 200	10, 100, 150, 200, 250
Min. no. of samples per leaf	2	None, 2, 3, 4, 5, 8, 10
Max. no. of leaf nodes	Not applied	None, 20, 40, 60, 80
Max. depth	None, 5, 10, 15, 20	None, 2, 5, 7, 8, 10, 12, 15, 25
Min. samples needed to split a node	2, 4, 6, 8	Not applied

Table A.1: Model parameters used in grid search for the random forest models in this project. Other values were set as the default values.

Parameter	Regression	Classification
l_1-ratios	0.001, 0.1, 0.5, 0.9, 0.99	0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9, 0.95, 0.99, 1

Table A.2: Model parameters used in grid search for the elastic net models in this project. Other values were set as the default values.

B

Appendix B

Table B.1: List of features mentioned in the report with explanations, abbreviations used and the number of subjects with a measure of the feature at baseline. If a variable name ends with IC, it mean missing indicator.

Abbreviation	**Name in database**	**Explanation**	**n**
ABETA40	ABETA40	Amyloid β 1-40 from CSF.	1210
ABETA42	ABETA42	Amyloid β 1-42 from CSF.	1210
ABETARatio	ABETARatio	$A\beta$-ratio. The ratio between $A\beta_{42}$ and $A\beta_{40}$ measured in CSF.	1210
ADAS11	ADAS11	The Alzheimer's Disease Assessment Scale–Cognitive Subscale, 11 item version. A cognitive test score.	2241
ADAS13	ADAS13	The Alzheimer's Disease Assessment Scale–Cognitive Subscale, 13 item version. A cognitive test score.	2229
ADASQ4	ADASQ4	ADAS Delayed Word Recall. A cognitive test score.	2246
Age	Age	Age of a subject.	2250
AMD	AMD_IPVDEEAFVIDFKPR	Peptidyl-glycine alpha-amidating monooxygenase in CSF.	287
APGEN2 2/3/4	APGEN2_2.0, 3.0, 4.0 and nan	APOE genotype - Allele 2.	2047
Apo H	Apolipoprotein H (Apo H) (ug/mL)	Apolipoprotein H in plasma.	566
APOB	APOB_SVSLPSLDPASAK	Apolipoprotein B-100 from CSF.	287

APOC-III	Apolipoprotein C-III (Apo C-III) (ug/mL)	A protein encoded by the APOC3 gene. Beleived to inhibit uptake of triglyceride-rich particles in the liver [23].	566
APOE4	APOE4	Apolipoprotein E gene with allele epsilon 4. A risk factor for AD which people can have zero, one or two copies of.	2051
AV45	AV45	Reference region - florbetapir mean of whole cerebellum. Regions defined by Freesurfer.	1050
BMP-6	Bone Morphogenetic Protein 6 (BMP-6) (ng/mL)	Bone morphogenetic protein 6. A protein that in humans is encoded by the BMP6 gene.	566
BTC	Betacellulin (BTC) (pg/mL)	A protein encoded by the BTC gene.	566
CA2D1	CA2D1_TASGVNQLVDIYEK	Voltage-dependent calcium channel subunit alpha-2/delta-1 in CSF.	287
CDRSB	CDRSB	Clinical Dementia Rating Scale–Sum of Boxes. A cognitive test score.	2253
CFAB	CFAB_YGLVTYATYPK	Complement factor B in CSF.	287
CH3L1	CH3L1_VTIDSSYDIAK	Chitinase-3-like protein 1. A protein encoded by the CHI3L1 gene.	287
CMGA	CMGA_SEALAVDGAGKPGAEEAQDPEGK	Chromogranin-A in CSF.	287
CNTP2	CNTP2_VDNAPDQQNSHPDLAQEEIR	Contactin-associated protein-like 2 in CSF.	287
DIGITSCOR	DIGITSCOR	Digit Symbol Substitution. A cognitive test score.	814
EcogSPLang	EcogSPLang	Measurement of Everyday Cognition (Ecog) Study partner report, Language factor. A cognitive test score.	1406

EcogSPMem	EcogSPMem	Measurement of Everyday Cognition (Ecog) Study partner report, Memory factor. A cognitive test score.	1404
EcogSPOrgan	EcogSPOrgan	Measurement of Everyday Cognition (Ecog) Study partner report, Organization domain score. A cognitive test score.	1348
EcogSPTotal	EcogSPTotal	Measurement of Everyday Cognition (Ecog) Study partner report. Total of scores. A cognitive test score.	1404
Entorhinal	Entorhinal	University of Claifornia, San Fransisco (UCSF) Entorhinal size in mm^3.	1468
FAQ	FAQ	Functional Activities Questionnaire. A cognitive test score.	2227
FDG	FDG	Fluorodeoxyglucose. A radiopharmaceutical used in PET scans. The uptake of FDG in the brain provides estimates of the cerebral metabolic rate of glucose (CMRglc).	1454
FP AB40	FREE_-PLASMA_-ABETA40	Free ABeta40 in plasma.	290
FP AB42	FREE_-PLASMA_-ABETA42	Free ABeta42 in plasma.	285
Fibrinogen	Fibrinogen (mg/mL)	Fibrinogen in plasma.	566
Final ASYN	FINAL_ASYN	Final reported concentration of alpha-synuclein.	368
Fusiform	Fusiform	University of Claifornia, San Fransisco (UCSF) Fusiform size in mm^3.	1468
Gender	GENDER	Gender of a patient.	2253
GLN	GLN	Glutamine, an α-amino acid that is used in the biosynthesis of proteins.	1638
GOLM1	GOLM1_QQL-QALSEPQPR	Golgi membrane protein 1 measured in CSF.	287
Hippocampus	Hippocampus	University of Claifornia, San Fransisco (UCSF) Hippocampus size in mm^3.	1492

ICV	ICV	Intracranial volume.	1725
KAIN	KAIN_LGFT-DLFSK	Kallistatin protein measured in CSF. It is encoded by the SERPINA4 gene.	287
L LDL FC	L_LDL_FC___	Free cholesterol to total lipids ratio in large LDL, i.e., low density lipoproteins.	1640
L HDL FC	L_HDL_FC___	Free cholesterol to total lipids ratio in large HDL, i.e., High density lipoproteins.	1641
LDELTOTAL	LDELTOTAL	Logical Memory - Delayed Recall. A cognitive test score.	2248
LRC4B	LRC4B_LTT-VPTQAFEY-LSK	Leucine-rich repeat-containing protein 4B in CSF.	287
M LDL P	M_LDL_P	Concentration of medium LDL particles, i.e., low density lipoproteins.	1641
M VLDL PL	M_VLDL_PL	Phospholipids in medium VLDL, i.e., very low density lipoproteins.	1641
MCP3	Monocyte Chemotactic Protein 3 (MCP-3) (pg/mL)	Monocyte Chemotactic Protein 3 in plasma (pg/mL).	566
MidTemp	MidTemp	University of Claifornia, San Fransisco (UCSF) Midtemp size in mm^3.	1468
MMP2	Matrix Metalloproteinase-2 (MMP-2) (ng/mL)	Matrix Metalloproteinase-2 in Plasma (ng/mL).	566
MMSE	MMSE	Mini Mental State Examination. A cognitive test score.	2253
MOCA	MOCA	Montreal Cognitive Assessment (MoCA) Test for Dementia. A cognitive test score.	1399
mPACCdigit	mPACCdigit	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Digit Symbol Substitution. A cognitive test score.	2249

mPACCtrailsB	mPACCtrailsB	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Trails B. A cognitive test score.	2249
MUC18	MUC18_-GATLALTQ-VTPQDER	Cell surface glycoprotein MUC18.	287
NGRN CV	NGRN_CV	Neurogranin coefficient of variation (%). Neurogranin is a calmodulin-binding protein expressed primarily in the brain.	144
PLASMA_NFL	PLASMA_NFL	Plasma neurofilament light (NFL).	1452
Plasma sample tag 0/1/nan	TAG_-PLASMA_-SAMPLE_0.0, 1.0 and nan	Plasma sample. Sample taken from plasma.	1641
PLDX1	PLDX1_LYG-PSEPHSR	Plexin domain-containing protein 1. A protein encoded by the PLXDC1 gene.	287
PPN	PPN_VHQSP-DGTLLIYNLR	Papilin protein in CSF.	287
Proinsulin Intact	Proinsulin- Intact (pM)	Proinsulin is the precursor of insulin during physiological insulin production. Intact proinsulin predicts progression of insulin resistance.	566
PTAU	PTAU	CSF PTAU. Phosphorylated tau protein.	1215
RAVLT I	RAVLT_-immediate	Rey's Auditory Verbal Learning Test (RAVLT) Immediate (sum of 5 trials). A cognitive test score.	2242
RAVLT F	RAVLT_-forgetting	Rey's Auditory Verbal Learning Test (RAVLT) Forgetting (trial 5 - delayed). A cognitive test score.	2241
RAVLT P F	RAVLT_-perc_forgetting	Rey's Auditory Verbal Learning Test (RAVLT) Percent Forgetting. A cognitive test score.	2235
S VLDL C	S_VLDL_C	Total cholesterol in small VLDL, i.e., very low density lipoproteins.	1641
SCG3	SCG3_ELSA-ERPLNEQIA-EAEEDK	Secretogranin-3 in CSF. A protein encoded by the SCG3 gene.	287

SNAP	SNAP	SNAP-25, Synaptosomal-Associated Protein, 25kDa is a t-SNARE protein encoded by the SNAP25.	146
SPRL1	SPRL1_HSASDDYFIPSQAFLEAER	SPARC-like protein 1. A protein encoded by the SPARCL1 gene.	287
TAU	TAU	Total-τ protein in CSF.	1215
TG PG	TG_PG	Ratio between triglycerides and phosphoglycerides.	1637
TLCA/CDCA	TLCA_CDCA	A ratio between taurolithocholic acid and chenodeoxycholic acid which are bile acids.	1669
TRABSCOR	TRABSCOR	Trail-making test B. A cognitive test score.	2193
UBB	UBB_TLSDYNIQK	Polyubiquitin-B in CSF. A protein encoded by the UBB gene.	287
VAL	VAL	Valine, an amino acid.	1640
XL HDL CE	XL_HDL_CE	Cholesterol esters in very large HDL, i.e., high density lipoproteins.	1641