Norwegian University
of Life Sciences

Master's Thesis 2020 30 ECTS
Faculty of Science and Technology

Multiblock-model Analysis of Multi-source Alzheimer's Disease Data

Jora Singh Randhawa
MSc. Environmental Physics and Renewable Energy

This page is intentionally left blank.

Preface

This thesis marks the end of a 5-year integrated master's program in Environmental Physics and Renewable Energy at the Faculty of Science and Technology (REALTEK) at the Norwegian University of Life Sciences (NMBU) in 2020.

First, I would thank my supervisors, Associate Professors Oliver Tomic and Kristian Hovde Liland at REALTEK, for the support, feedback and guidance during this long process of writing this thesis. I would also thank Dr. Inge Groote and Dr. Per Selnes at the Computational Radiology and Artificial Intelligence (CRAI) research group at Oslo University Hospital for the guidance, discussions, and for providing the data.

Furthermore, I would like to thank my family and friends for the unconditional support. And a special thanks to my brother, Dr. Partap Singh, for guidance in the medical field.

Oslo, 15th December 2020

Jora Singh Randhawa

Abstract

Alzheimer's Disease (AD) is the most common cause of dementia in the world. It is a disorder that causes brain cells to degenerate and eventually dies, which causes a continuous decline in memory, cognitive abilities and social skills. As the disease develops, a person's ability to function and carry out daily tasks will eventually be impossible. There are currently no treatments that cure AD, making people affected by this disease dependent on others for assistance.

Detecting AD in the early stages will help slow down the disease's development and improve life quality for people affected. Early initiatives will allow patients to live with fewer health problems for a more extended period by changing their lifestyle.

This Master's thesis explored the usefulness of applying machine learning methods and data analytics to detect important risk factors for AD. Methods such as Partial Least Squares (PLS), Principal Component Analysis (PCA), feature importance permutation, and Sequential and Orthogonalized PLS (SO-PLS) were utilized to find relevant features and their importance. The measurement for AD was cerebrospinal fluid amyloid-beta (CSF betaA) in the spinal fluid and was used as the target with the supervised method used in this thesis.

The model developed to detect risk factors for AD accomplished an explained variance of 22.89 %. Important factors from the model were the Apolipoprotein E4 / E4, aggregated white matter hyperintensities (WMHs), aggregated lesion in the brain and lesion at the second layer in the parietal lobe.

Evaluating the results indicates that the model encountered insufficient data and block separation, which generated a poor performing model. The results indicate that no definitive risk factors can be identified as to what causes AD. The methods and the data still have a potential for improvement and further work.

Sammendrag

Alzheimers sykdom (AD) er den vanligste årsaken til demens i verden. Det er en lidelse som gjør at hjerneceller degenererer og til slutt dør, noe som fører til en kontinuerlig fall i hukommelse, kognitive evner og sosiale ferdigheter. Etter hvert som sykdommen utvikler seg, vil en persons evne til å fungere og utføre daglige oppgaver til slutt være umulig. Det er ingen behandlinger som kurerer AD, noe som gjør at folk som er rammet av sykdommen er avhengige av andre for hjelp.

Ved å oppdage AD i tidlig stadium kan man bidra til å bremse utviklingen av sykdommen og forbedre livskvaliteten for de berørte. Tidlige tiltak vil gi pasienter en mulighet til å leve med mindre helseplager i en lengre periode ved livsstilsendringer.

Denne masteroppgaven utforsket nytten av å anvende maskinlæringsmetoder og dataanalyse til å oppdage viktige risikofaktorer for AD. Metoder som Partial Least Squares (PLS), Principal Component Analysis (PCA), feature importance permutation, og Sequential and Orthogonalized PLS (SO-PLS) ble brukt for å finne relevante funksjoner og deres betydning. Målingen av AD var cerebrospinalvæske amyloid-beta (CSF betaA) i ryggmargsvæsken.

Modellen som ble utviklet for å oppdage risikofaktorer oppnådde en forklart varians på 22,89 %. Viktige risikofaktorer utarbeidet fra modellen var Apolipoprotein E4 / E4, aggregert hvit substans hyperintensiteter (WMH), aggregerte lesjoner i hjernen og lesjon ved det andre laget av parietallappen.

En evaluering av resultatene indikerer at modellen hadde utilfredsstillende data og en svak blokkinndeling, noe som førte til en svak modell. Utfallet av resultatene indikerer at ingen definitive risikofaktorer kan identifiseres for hva som forårsaker AD. Metodene brukt og innhentet data har fremdeles et potensial for forbedring og videre arbeid.

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Meaning
AD	Alzheimer's disease
CV	Cross-validation
IG	Information gain
KNN	K-Nearest Neighbors
MFA	Multiple factor Analysis
MRI	Magnetic resonance imaging
OLS	Ordinary least-squares
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Least Square
PLS-SVD	Partial Least Square-Singular Value Decomposition
RMSEP	Root Mean Square Error of Prediction
SGD	Stochastic Gradient Descent
SO-PLS	Sequential and Orthogonalised Partial Least Square
SSR	sum of squared of the residuals
STD	Standard Deviation
SVM	Support vector machines

Chapter 1

Introduction

1.1 Background

Dementia is a significant cause of disability among older adults worldwide. It affects the memory, cognitive abilities and behaviour, and will eventually affect one's daily activities. The effects of dementia is a significant cause of disability and dependency among older adults worldwide, which leads to severe impact on peoples families, career and communities alongside the individuals [1]. The root cause of dementia is not known, but several risk factors are known, such as ageing, inactivity, obesity, harmful use of alcohol, tobacco use and diabetes. However, there is no assurance that preventing these known factors will have an effect on every individual [2].

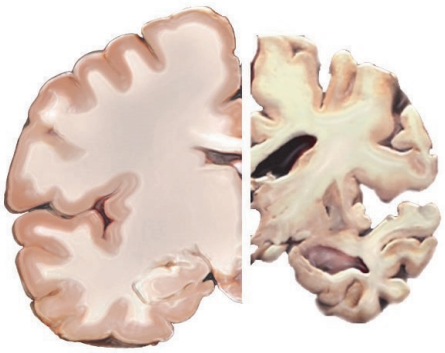


Figure 1.1: The figure shows a comparison of a healthy brain and a brain with severe Alzheimer's Disease [3].

Dementia is a collective term for several diseases; the most common one is Alzheimer's disease (AD). This disease affects about 3 percent of people over 65 years of age and about 12-15 percent of people over 80 years of age. AD starts in the brain several years before detecting any form of symptoms or signs. Because of this gradual development, it is hard to identify the disease in the early stages [4]. The risk factor of age plays an essential role in developing AD, as the distribution shows in Figure 1.2, high age, or more precise, the brain's age increases the probability of AD or developing AD [5].

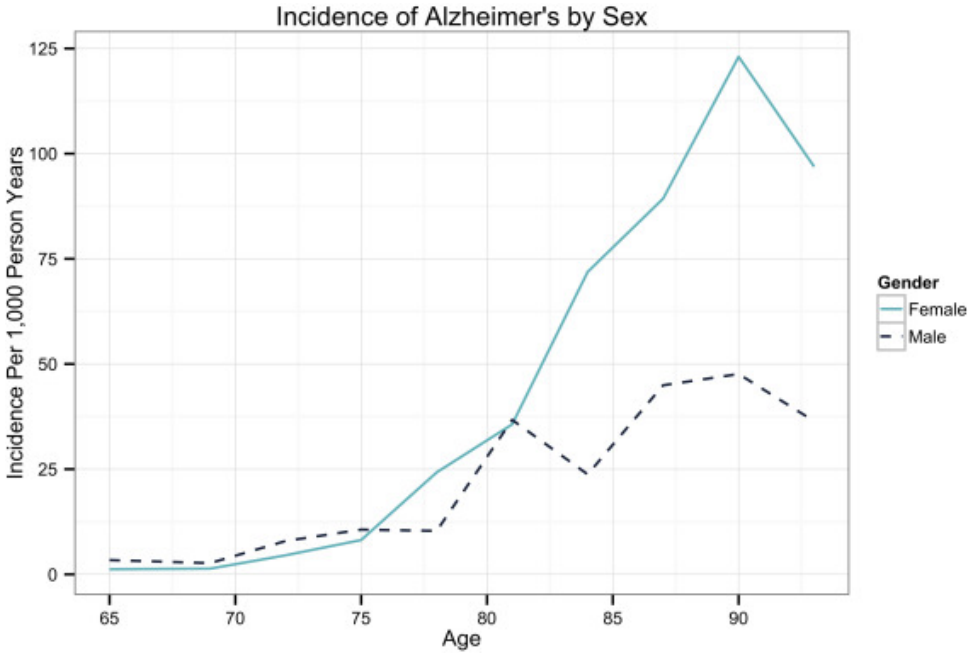


Figure 1.2: Age distribution of AD with sex-specific incidences per 1000 person years [5], which indicates that age is significant for developing AD.

Some clinical criteria for dementia syndrome and AD must be fulfilled to determine if a patient has AD. The criteria for AD are characterized by amyloid plaques and loss of neurons in the brain [6]. The criteria for dementia syndrome [7], according to ICD-10 (10th revision of the International Statistical Classification of Diseases and Related Health Problems) is as follows;

1. Significantly worse memory than before in life, especially for events in the recent past.

2. At least one other cognitive function decreased related to previous ones, such as logical reasoning and linguistic communication ability.

3. Reduced ability to function in daily life.
4. Changed behaviour, such as more passive, effortless or annoyed.
5. Symptoms described in 1 to 4 must be persisted for at least six months.
6. Still remain normal consciousness.

One of the most significant risk factors for AD is genetics, more precisely the ApoE-ϵ4 allele. Apolipoprotein E (ApoE) regulates lipid homeostasis by moderating fatty acid and lipid transport from one cell type to another [8]. The ApoE-ϵ4 allele has been involved in several diseases, including AD, such as HIV [9] and much recently COVID-19 [10]. The frequency of this allele compared to the two other polymorphic alleles, ϵ2 and ϵ3, is 13.7% compared to 8.4% and 77.9%, respectively. However, the frequency of the ϵ4 allele is significantly increased for patients with AD [8]. Some studies also show that synergies with ϵ4 allele and other vascular diseases [8][11][12].

To determine that a patient is developing AD or has AD, measurements of the accumulation of the protein fragment beta-amyloid (betaA) plaques outside neurons and measurements of the accumulation of tau inside the neurons are two essential changes in the brain for the decision-making [4][6]. Since these values are usually difficult to measure, because of the complexity of measurements in the brain, other methods must be used. Such methods are cognitive tests, such as mini-mental status evaluation (MMSE), Alzheimer's Questionnaire (AQ) [13], trail making test A (TMT-A) and trail making test B (TMT-B) [14]. These tests support medical experts in the diagnosis of AD and other diseases.

For the actual measurement of betaA in the brain, some other method must be used. Cerebrospinal fluid (CSF) betaA are used as a biomarker for AD [15]. CSF betaA is measured in the spinal fluid and are negative correlated with betaA. Low accumulation of CSF betaA in the spinal fluid indicates high accumulation if betaA in the brain.

1.2 Problem statement

Except for some known modifiable risk factors of AD [16], there exist no other documented modifiable risk factors for AD. The etiological risk factors, other than ageing and genetic proneness, remain to be determined [17]. There exist, therefore, limited information about the cause and prevention of AD.

Machine learning and data analysis have, in recent years, been used in the research of AD and other types of dementia. These fields have opened up a new way of analyzing complicated and large datasets, with the purpose of assisting and improving medical experts in their assessment. Such large datasets make it possible to study a larger amount of factors that may contribute to better understand AD in a systematic manner. This thesis aims to use principal component analysis (PCA) [18], partial least squares (PLS), and the multi-block regression method, sequential and orthogonalized partial least square (SO-PLS) [19] to determine which factors or variables improve the assessment of patients with AD.

The thesis will focus on the accumulation of CSF beta-amyloid as the measurement of AD, where data and variables are provided from Computational Radiology and Artificial Intelligence (CRAI) seated in Oslo University Hospital. The thesis tries to find and understand features or risk factors that indicates high values of beta-amyloid.

1.3 Previous work

SO-PLS has been mainly used in the chemometrics field. In this thesis, SO-PLS will be used as a multiblock method to analyse heterogeneous data from various sources to better understand AD and different levels of beta-amyloid. The SO-PLS method can be beneficiary for the analysis because it takes different sources or measurements into account when applying the model.

Relevant information and background of SO-PLS regression are explained in "Path modelling by sequential PLS regression"[20] and "SO-PLS as an exploratory tool for path modelling" [21]. These papers explains the methodology or path modelling and the proper usage of SO-PLS.

Recent studies that apply SO-PLS, such as "SO-PLS as an alternative approach for handling multi-dimensionality in modelling different aspects of consumer expectations" [22], has also been studied for this thesis.

1.4 Structure of thesis

This thesis starts with the theory behind machine learning, more specifically, the theory behind PCA and PLS in chapter 2. In chapter 3, the datasets are studied

and prepared, and chapter 4 the methodology of SO-PLS is described. Chapter 5 covers the results, which are discussed in detail in chapter 6. The results of this thesis are summarised in chapter 7.

Chapter 2

Theory

Machine learning is a subset of artificial intelligence (AI), as shown in Figure 2.1, which uses some statistical models to perform a task that predicts the outcome by recognizing patterns and dependencies. Machine learning builds models that predict the outcome by analyzing data that would else be tedious and difficult for humans to do [23]. Such a process can be useful for understanding the behaviour and properties of a known system.

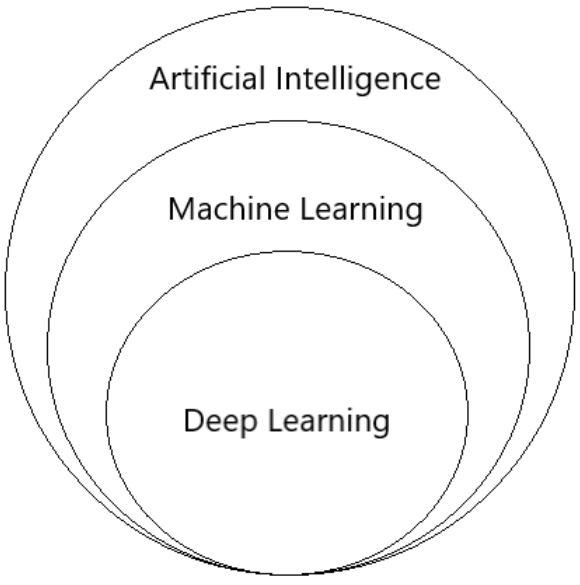


Figure 2.1: The figure shows the placement of machine learning compared to artificial intelligence and deep learning.

There are three paradigms of machine learning; supervised learning, unsupervised learning, and reinforced learning. Supervised learning is using labelled data to train a model that predicts the future outcome of unlabelled data. Such learning models can be separated into two classes that are, classification models and regression models. In classification, models are trained to classify a given set of classes, or categories, to assign new data points to given groups. The model is built based on given data points that are fitted based on a decision boundary which assign the data points to their given category. Regression models, such as Partial Least Square (PLS) regression, use continuous response values to assign rather than given classes or categories [23].

Unsupervised learning methods analyses a data set without a response variable. The goal is to showcase the underlying structure or distribution of the data set for understanding more about the data and its underlying systematic variation [23]. Such analyzing models is Principal Component Analysis (PCA), which finds components that describe systematic variation in the data.

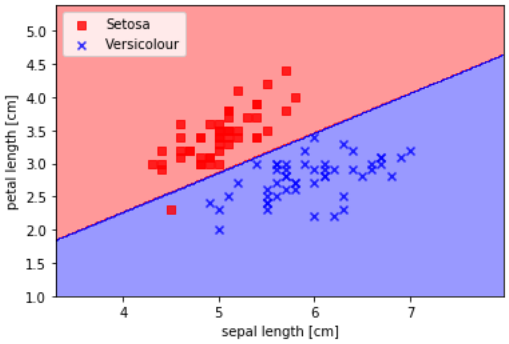


Figure 2.2: The figure shows a classification model model, Perceptron, with the Iris data set [24]. It contains two classes, *Versicolor* and *Setosa*, and a decision boundary that separates the classes.

2.1 Least Squares

Least-squares is a method for performing linear regression. There exist two categories where least-square problems commence; ordinary least-squares (OLS) and nonlinear least squares. This thesis will investigate a closed-form solution where the ordinary least-squares is applicable, linearity is maintained. OLS minimizes the sum of square of the residuals, which leads to the estimated value of the unknown parameters α and β [25], where α are the bias and β is the regression

coefficient.

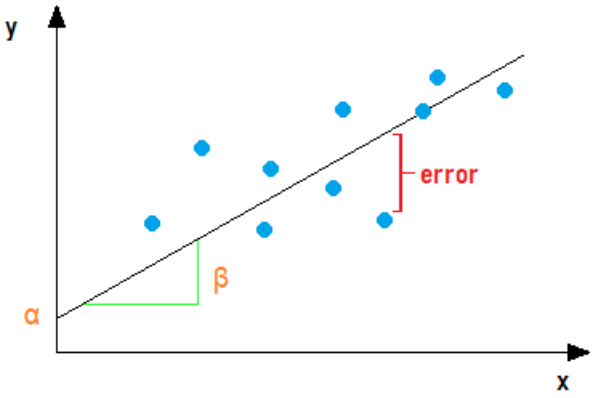


Figure 2.3: The figure shows a simple linear regression model, where the goal is to minimize the sum squared of the residuals. The parameters α and β are the estimated values, $\hat{\alpha}$ and $\hat{\beta}$, shown in formula 2.4. The error illustrated in the figure indicates one single error, ε_i, shown in formula 2.1 and 2.2. *source:[25][26].*

Suppose the data consists of k observations, y_k and x_k, where y and x represent the response and the variables, respectively. Which is represented in formula 2.1 and as vector form in formula 2.2, where β is a $\{k \times 1\}$ vector and ε_i is the error term.

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{2.1}$$

$$y_i = \alpha + x_i^T \beta + \varepsilon_i \tag{2.2}$$

As mentioned for OLS, the goal is to minimize the sum of squared of the residuals (SSR). This means that the estimated values for α and β, $\hat{\alpha}$ and $\hat{\beta}$, obligate to provide the lowest value for SSR [25]. SSR is calculated in formula 2.3, where b is a estimated value for the parameter $\hat{\beta}$ and a is a estimated value for the parameter $\hat{\alpha}$. The value of b and a that minimizes SSR denotes the value for $\hat{\alpha}$ and $\hat{\beta}$.

$$SSR(a, b) = \sum_{i=1}^{k} (y_i - x_i^{\mathrm{T}} b - a)^2 \tag{2.3}$$

Furthermore, calculating formula 2.3 for α and β, separately, gives us the solution to the OLS model that minimizes the squared errors, shown in formula 2.4 and visualized in Figure 2.3.

$$y = \hat{\alpha} + \hat{\beta}x \tag{2.4}$$

2.2 Partial Least Square Regression (PLSR)

2.2.1 Background

Partial Least Squares Regression (PLSR) was developed by Wold et al. [27] and is a method for the linear modelling of the relation between the variables X and the response Y [28], that can be used for both univariate and multivariate regression. The PLSR-algorithm tries to find the components that maximize the covariance between the response and explanatory variables with the intention to capture most of the information in X that is useful for predicting Y while reducing the dimension of the model [29].

PLS is a latent-variable based method, which means that the model's primary goal is to describe the observed variables in terms of the latent variables. It constructs a new set of variables (latent variables) from the linear combination of predictor variables, X. This goal is achieved and explained by projecting the variables X and the response Y into a new space matrix. However, different PLS algorithms may achieve their goal differently.

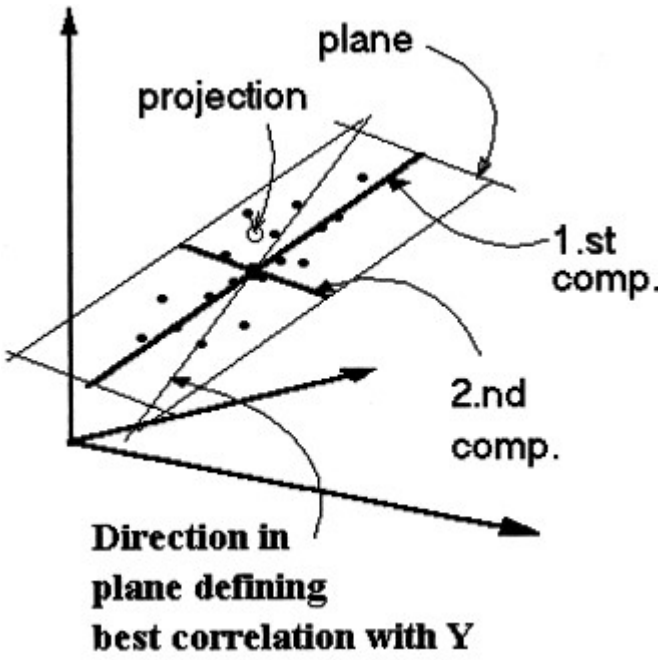


Figure 2.4: An overview and a geometric representation of PLS regression in 3 dimensions. The figure shows two principals components calculated with PLS and the direction in the plane that best defines the correlation between the response, Y, and the variables, X. *source:[26].*

There exist different variants of PLS that has its origin from Wold's work [30]. The original work from Wold can be divided into two modes; A and B. Wold's method includes three key parts; it is a class of algorithms that contains arbitrary number of blocks of indicators with their latent variables, an arbitrary linear relation between the latent variables, and the computation that are separated into modes. The difference in the modes are the computation that are interpreted differently [31]. Furthermore, several other mode-A algorithms have been developed, such as PLS-SVD, PLS1 and PLS2, shown in figure 2.5.

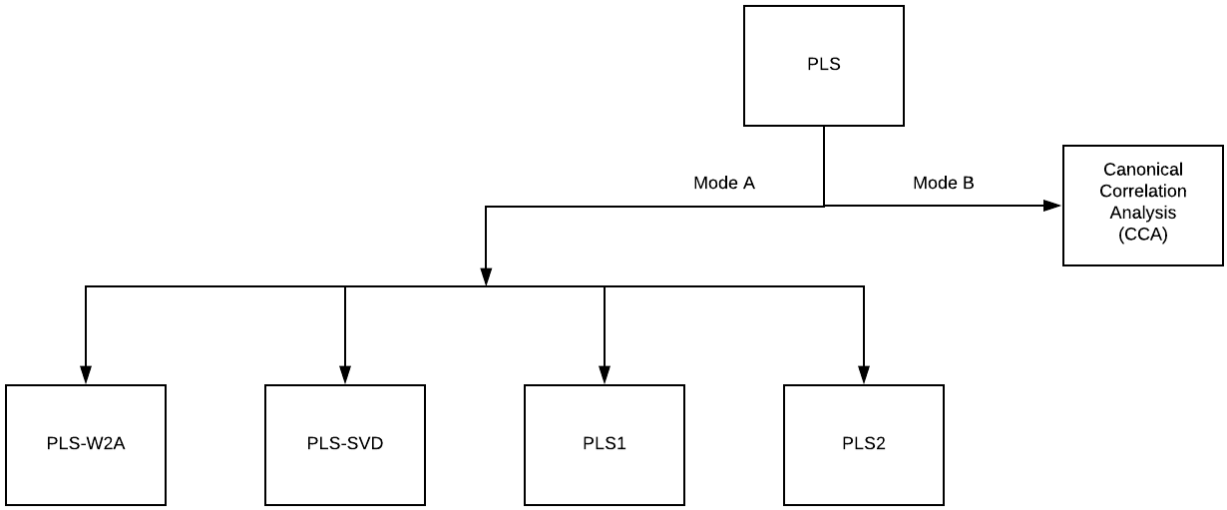


Figure 2.5: An overview of PLS algorithms originated from the original work of Wold [30][31]. The original work of Wold is Wold's Two-Block Mode A PLS (PLS-W2A) and Canonical Correlation Analysis (CCA). Even though mode A has mainly been associated with PLS, the CCA belongs to the class of PLS [31].

Furthermore, it can be shown that results from multi-block PLS methods can be calculated from two-block PLS methods if the same scaling of variables is applied [32]. This thesis will concentrate on multi-block PLS methods, rather than two-block PLS. Such methods, widely used in chemometrics [31], are uniresponse PLS (PLS1) and multiresponse PLS (PLS2). The difference between PLS1 and PLS2 is the response variable. PLS1 only considers one single response column at a time, while PLS2 requires multiple response columns.

2.2.2 General model

The general underlying model for multi-block or multivariate PLS is shown in formula 2.5 and formula 2.6, where X is a matrix of predictor variables with N observations and K variables ($N \times K$), and Y is a matrix response variable with N observations and M variables ($N \times M$).

$$X = TP^T + E, \qquad E \sim \mathcal{N}(\mu, \sigma^2) \tag{2.5}$$

$$Y = UQ^T + F, \qquad F \sim \mathcal{N}(\mu, \sigma^2) \tag{2.6}$$

T is a $N \times l$ matrix projection of X (the X-scores), and U is a $N \times l$ matrix projection of Y (the Y-scores). P and Q are $K \times l$ and $M \times l$ orthogonal loading matrices, respectively. Furthermore, matrices E and F are the error terms, assumed to be independent and identically distributed normal variables [33].

The goal of the PLS explained through formula 2.5 and formula 2.6 is to maximize the covariance between T and U while minimizing the norm of F [33]. The solution to the mentioned issue is to find the optimum number of principal components, shown in figure 2.4, that gives the lowest value of Root Mean Squared Error of Predictions (RMSEP). This is achieved by empirically determine the RMSEP by **cross-validation**, which will be further elaborated in chapter 2.4.2. RMSEP is defined in formula 2.7 and visualized in a RMSEP plot in figure 2.6. In addition to RMSEP, the Predictive Error Sum of Squares (PRESS) is also used in achieving the optimal number of components.

$$RMSEP = \sqrt{\frac{PRESS}{N}} = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}} \tag{2.7}$$

Where N is the number of predicted observations, and $\hat{y}_i$ is predicted values of the variable y_i. RMSEP and PRESS is useful metrics for estimating performance because large errors will be enhanced. This is because significant errors have a more substantial impact on the score than small ones, which leads to the certainty that the lower the RMSEP value, the higher the predictive ability of the model. The disadvantage of RMSEP is the high number of calculations needed to obtain the RMSEP value necessary for achieving the right amount of principal components.

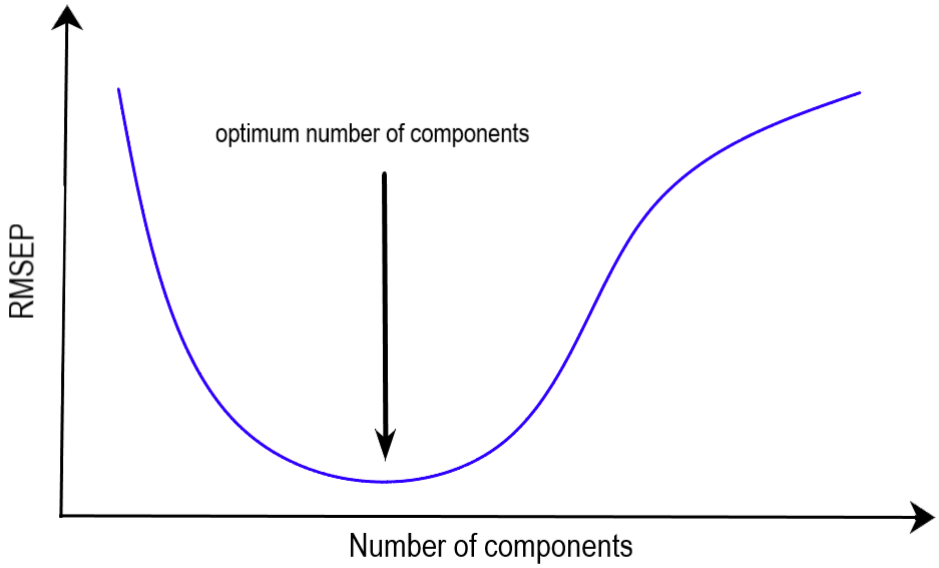


Figure 2.6: A plot of RMSEP calculations with the best model shown as the optimal number of components. It should be noted that number of components is a discrete variable.

2.2.3 Prediction

The prediction of new response values is managed differently depending on PLS variant. The focus in this thesis will be on **uniresponse** PLS and **multiresponse** PLS.

For PLS1 the computation of the estimated values, $\hat{\beta}$, with k principal components can be written as [34];

$$\hat{\beta} = \hat{W}(\hat{P}^T\hat{W})^{-1}\hat{q} \tag{2.8}$$

Where $\hat{W}$ is the loading weights, $\hat{W} = [\hat{w_1}...\hat{w_k}]$, for each principal component scaled to length 1. $\hat{P}$ is the estimated X-loadings for each principal component, $\hat{P} = [\hat{p_1}...\hat{p_k}]$, and $\hat{q} = [\hat{q_1}...\hat{q_k}]$ is the estimated Y-loadings. The estimated X-loadings and Y-loadings is derived from formula 2.9 and formula 2.10 respectively.

$$\hat{p}_k = \frac{X_{k-1}^T \hat{t}_k}{\hat{t}_k^T \hat{t}_k} \tag{2.9}$$

$$\hat{q}_k = \frac{y_{k-1}^T \hat{t}_k}{\hat{t}_k^T \hat{t}_k} \tag{2.10}$$

Where $\hat{t}$ represent the estimated scores, $\hat{t} = X_{k-1}\hat{w}_k$. The estimated loadings weights can further be written as $\hat{w}_k = X_{k-1}^T y_{k-1}$.

For the multiresponse PLS2, we interpret y and q as matrices rather than vectors. Instead of using y_{k-1}, the new vector $\hat{u_k}$ is introduced. The following three steps are repeated until the estimated scores $\hat{t_k}$ converges [21]:

1. Calculate the loadings weights, $\hat{w_k} = X_{k-1}^t \hat{u_k}$, and scale the weights to length 1.
2. Estimate the scores, X-loadings and Y-loadings the same as for uniresponse PLS.
3. Check if the scores, $\hat{t_k}$, has converged. If not, estimate the new vector $\hat{u}_k = Y_{k-1}\, \hat{q}_k(\hat{q}_k^t \hat{q}_k)^{-1}$.

When $\hat{t}_k$ converges, the same procedure for estimating β as for uniresponse follows, the formula 2.8 is used.

2.3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is an unsupervised method that provides a dimensionality reduction from the original dataset. Unsupervised methods have some benefits, such as not requiring or relying on a response dataset for analyzing the data. PCA is used as a tool in exploratory data analysis as a technique to find the main characteristics in the data. This could be useful for finding hidden structures in a complex and large dataset, and to perform feature extraction and feature elimination.

The main goal of PCA is to project high-dimensional data space onto a two-dimensional data space in such a way that features of the dataset will be separable and visible. By projecting the data with its principal components, shown in Figure 2.7, will satisfy the main goal of PCA.

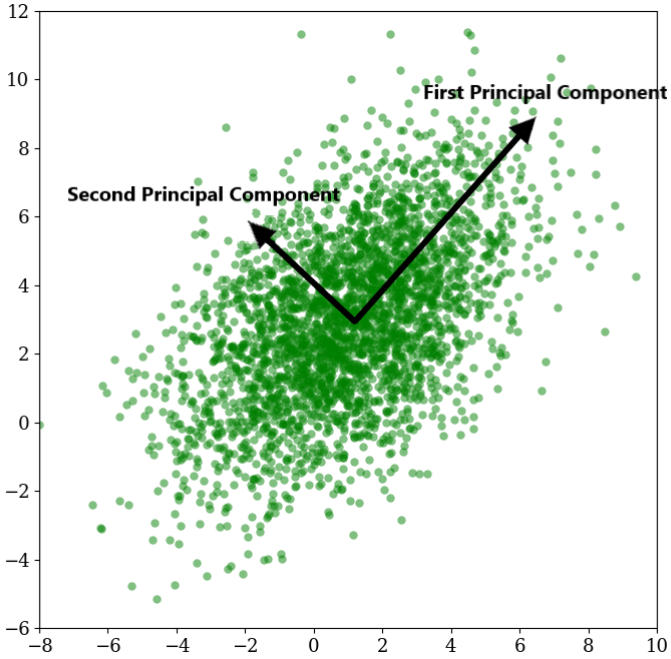


Figure 2.7: An overview and a geometric representation of PCA of a multivariate Gaussian distribution displayed in 2 dimensions. The vectors shown are the eigenvectors of the covariance matrix, the principal components.

PCA is accomplished by doing the following steps:

1. Calculate the data **correlation matrix** from the original dataset.
2. Carry out an **eigenvalue decomposition** on the correlation matrix.

The correlation matrix is found by calculating the correlation coefficient between each variable in the dataset, X. Formula 2.13 shows the equation for calculating the correlation coefficient between two arbitrary variables in X, X_i and X_j. Where *corr* and *cov* indicate correlation and covariance, respectively.

$$\rho_{X_i,X_j} = \operatorname{corr}(X_i, X_j) = \frac{\operatorname{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}} = \frac{\mathrm{E}[(X_i - \mu_{X_i})(X_j - \mu_{X_j})]}{\sigma_{X_i}\sigma_{X_j}} \tag{2.11}$$

σ_X is the standard deviation, and μ_X is the expected value, while E is the expected value operator. By calculating the correlation coefficient between each variable, it is possible to construct a correlation matrix. This is shown in formula 2.12, where n indicates the total number of variables.

$$\text{corr(X)} = \begin{bmatrix} 1 & \rho_{X_1,X_2} & \cdots & \rho_{X_1,X_n} \\ \rho_{X_2,X_1} & 1 & \cdots & \rho_{X_2,X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_1,X_n} & \rho_{X_n,X_2} & \cdots & \rho_{X_n,X_n} \end{bmatrix} \tag{2.12}$$

Since the correlation between X_i and X_j is the same as the correlation between X_j and X_i, the correlation matrix is a $[n \times n]$ symmetric matrix.

The next and final step of PCA is the eigenvalue decomposition of the correlation matrix. For simplification, the correlation matrix will hereafter be denoted as X rather than $corr(X)$. By decomposing the correlation matrix into its eigenvectors and eigenvalues, and sorting it in decreasing order, the final step of PCA is completed. The first eigenvector, in the eigenvector matrix, will then be the first principal component (PC), and the second eigenvector will be the second PC and so on. Formula 2.13 shows a decomposition of a matrix X into two matrices V and U.

$$X = U * V^T \tag{2.13}$$

Where V and U are referred to as **loadings** matrix and **scores** matrix, respectively. Loadings matrix can be understood as the weight of each variable when calculating the PCs; thus, a loadings plot showcases the variables with the given number of PCs as axis, shown in Figure 2.8. The scores matrix contains the original data rotated in a coordinate system with the given PCs as axis, shown in Figure 2.8.

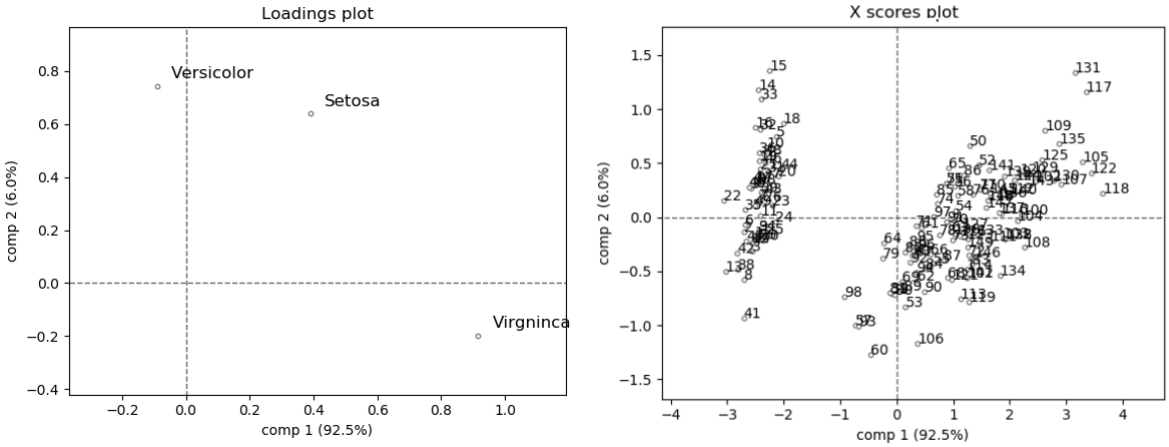


Figure 2.8: Loading plot (left plot) and score plot (right plot) of the Iris dataset [24] after performing PCA.

2.4 Model validation

The quality of a dataset is determined by how well it represent its intended use. Moreover, a dataset is considered of high quality if it satisfies the requirement of its intended usage and provides consistency. To ensure that a dataset is providing consistency throughout the analysis, some procedures must be followed. This section will discuss some practices that provide such data quality.

2.4.1 Test set

Standard practice for data analysis is splitting the data into two sets; **training set** and **test set**, as shown in Figure 2.9. The training set is used in the training the model, while the test set is used for validation and checking for consistency in the model. The test set's purpose is to verify the model's purpose when exposed to new data. This is a key criteria for indicating a good predicting model.

However, a model is considered **overfitted** when it has a good performance on the training set, but a poor performance on the test set. An overfitted model is usually detected by the test set when there is a poor performance on the test set, while the overall performance is adequate. We seek to find a well performing model that has similar performance on the test set and the training set.

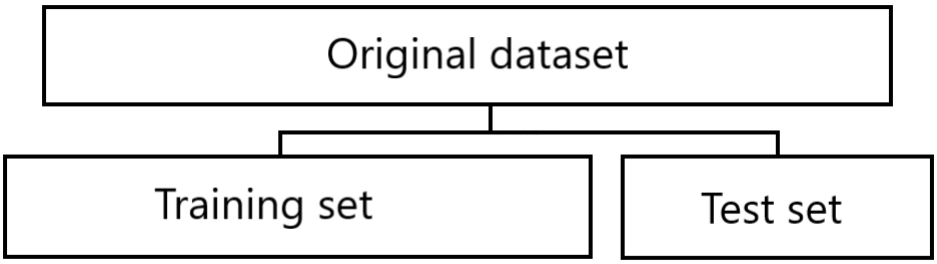


Figure 2.9: The original dataset split into a training set and a test set.

The size of the test set depends on the dataset available. The distribution of the samples is vital for a well-distributed test set and training set. A dataset that represents a real-world case and is sufficiently large will have a training set and test set that gives a decent model.

2.4.2 Cross-Validation (CV)

Cross-validation (CV) is a model validation technique for evaluating how well the model will **generalize** when exposed to a new independent dataset. The training set is usually divided into a new training set and a validation set, where the model performs the analysis on the training set and then validates or tests the result in the validation set.

K-fold cross-validation is a CV method that divides the training set randomly into K equally sized samples. Of the K subsamples, one single sample is used as the validation data, while *K-1* samples are used as training data. This process is repeated K times with each sample used as the validation set once. An example of this is shown in Figure 2.10. The result of the models in Figure 2.10 can be averaged into one single estimate.

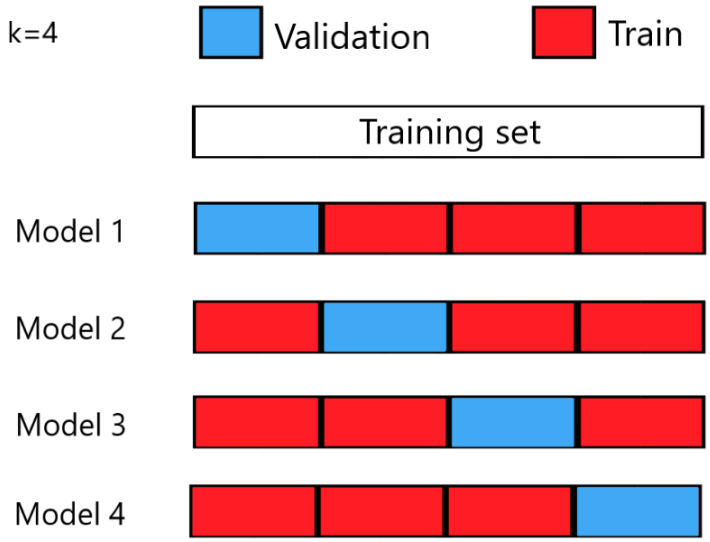


Figure 2.10: A K-fold cross-validation with the training dataset is split into new training sets and test sets. In this figure, k equals 4.

2.5 Classification methods

This subsection gives a brief explanation of different machine learning classification methods used in the computation of feature importance process (section 2.6) in this paper. In general, classification techniques are a supervised machine learning process used in the prediction of groups or, more specifically, classes from new observations. These classification problems and classes can be binary problems, such as decisions rated as success or failure, or if a patient has Alzheimer's disease or not. Other classification problems, called multi-class classifiers, are divided into more classes. An example is problems where the classification method tries to predict different age groups from some pre-defined data. These age groups or classes can be divided into decades, such as people in the ages of 0 to 9 is one class and so on. This thesis will further empathize with binary classification problems.

Before explaining different classification methods, some basic understanding of machine learning algorithms for classification and terminology needs to be established. The figure 2.11 shows the neural network perceptron, with input values, weights, net input function, activation function, output and error values.

The inputs that are shown in figure 2.11 refers to the data used in the training of the model. Each row in the dataset is represented as a vector $[x_1, x_2, ..., x_m]$ where

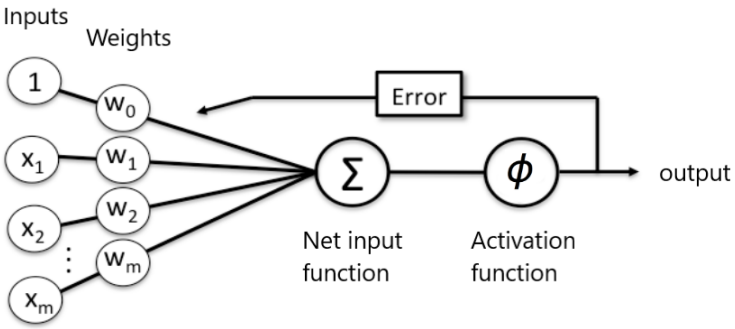


Figure 2.11: An illustration of a standard neural network algorithm with input values, weights, net input function, activation function, output and error values.

m is the number of columns in the dataset. These vectors are used to train the given model and later on update the weights. **The weights** could be of several layers, but for simplification there is one layer, as shown in figure 2.11. The weights are values that are associated with each input value which tells the importance of each input value. If an arbitrary weight contains a high value, the input values associated with it are of high importance and is a key feature to the final model. Figure 2.11 does not show the bias which is exclusive to each weight layer. The bias can be referred to as the y-intercept for a linear model (see section 2.1).

The **net input function**, also referred to as the summation function, is the summation of weight with its given input shown in Formula 2.14, where b_i is the bias from each layer.

$$sum = \sum_{i=1}^{m}(w_i * x_i) + b_i \tag{2.14}$$

The **activation function**, which is also shown in figure 2.11, transform the net input. The function decide how each input should be weighted. If the activation function is a linear function, such as $y = x$, then each input is weighted equally to the error and to the output.

The **error** term, along with the **cost function**, is used to update the weights. The updating of weights applies after each row or sample of data is iterated through the neural network, and is shown in Formula 2.15 and 2.16.

$$w_i := w_i + \Delta w_i \tag{2.15}$$

$$\Delta w_i = \eta(y^{(j)} - \hat{y}^{(j)})x_i^{(j)} \tag{2.16}$$

Where i is each weight and j is each training sample. y and $\hat{y}$ represent the true class label and the estimated class label respectively. The differentiation between the true class label and the estimated is known as the error. The learning rate, η, determines how much the updating should affect the initial weights.

Finally the output from the classification model could either be the output from the activation function or it could binary classified through a **threshold function**. A threshold function is shown in Formula 2.17, where input values bigger than zero gives an output one and otherwise minus one.

$$T(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ -1 & otherwise \end{cases} \tag{2.17}$$

2.5.1 Logistic Regression

Logistic regression is indeed a classification algorithm, which predicts the probabilities of each class. It is named after the logistic function (*logit*-function) which is shown in Formula 2.18. This function is explained as the logarithm of the odds, where p is the probability of the positive events or the preferred outcome [23].

$$logit(p) = \log(\frac{p}{1-p}) \tag{2.18}$$

The actual interest in this function is to predict the probability of each sample belonging to a particular class, which can be expressed by the inverse of the *logit* function, shown in Formula 2.19. This function is known as the logistic sigmoid function [23].

$$\phi(z) = \frac{1}{1-e^{-z}} \tag{2.19}$$

Formula 2.19 shows the **activation function** $\phi(z)$ for the logistic regression model. This activation function is also shown in figure 2.12. The mentioned activation function defines logistic regression and is unique for this model.

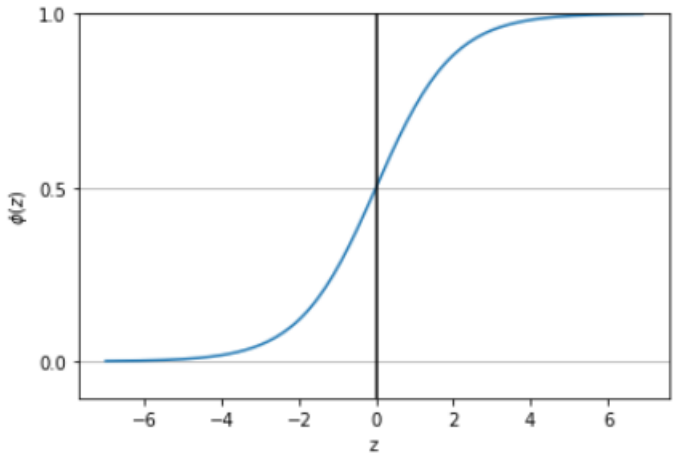


Figure 2.12: An illustration of the sigmoid function, also mentioned in Formula 2.19, with z as the x-axis and activation $\phi(z)$ as the y-axis.

2.5.2 Support Vector Machine (SVM)

Support vector networks [35] or support vector machines (SVMs) are both used in classification and regression analysis and is a linear classifier. A linear classifier labels data into classes based on a linear combination of the input values given to the model. SVM models use multiple hyperplanes to achieve a separation between different classes. The amount of features, k, given to the model shows how many hyperplanes are needed for the model, which are $k - 1$. SVM are therefore a good application for classifying linear separable data.

The objective of SVMs is to maximize the margin between the separating hyperplane and the closest classified data points. In some perspective, the objective for other models, such as the linear regression, is to minimize the error between predicted values and true values. The margin and SVM are shown in figure 2.13.

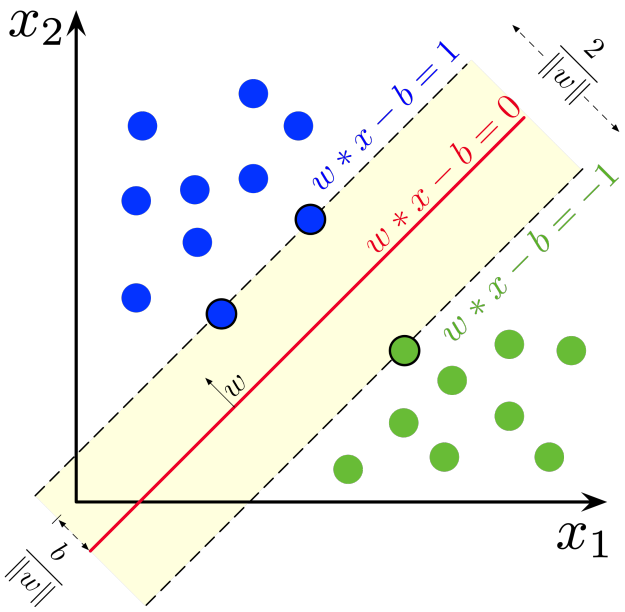


Figure 2.13: The figure illustrates SVM with its margin. SVM calculates the margin of the closest points to the hyperplane. In this illustration, the data is represented by two vectors x_1 and x_2, and b and $\vec{w}$ represent the intercept and margin vector respectively. The vector $\vec{w}$ can also be explained as the normal vector to the hyperplane. *source: [36]*

2.5.3 Decision Tree

Decision trees are used in machine learning as a predictive classification model. In general, decision trees is a tool used to map possible outcomes in a tree-like manner. Such tree are shown in figure 2.14 where two decisions or choices are shown.

Decision trees can be interpreted as a model that tries to break down the dataset by asking a series of questions. When applying this classifier, the splitting of the data is applied in such a manner that it maximizes the **information gain (IG)**. The information gain is the measured objective that we base the construction of the tree upon.

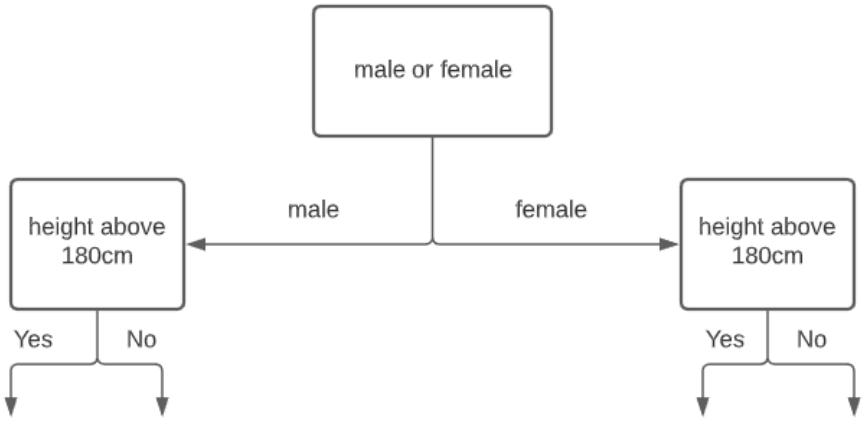


Figure 2.14: A simple example of a decision tree where the labels or features, sex and height, are displayed.

2.5.4 Random Forest

Random forest classifier can be regarded as an ensemble of decision trees. It is an average over multiple decision trees, to build a more robust and generalized model. It can be summarized in four steps [23]:

1. Construct a random sample, n.

2. Construct a decision tree from the samples and for each node; select, k random features and split the node which gives the highest information gain.

3. Repeat step 1 and 2 k times.

4. Accumulate the prediction of the trees and assign class labels by **majority vote**.

Majority voting is assigning a class label to a given sample based on that it received the majority of the votes from the predictive models. In this case the class is assign to the sample if that given class label received more than 50 percent of the votes from the decision trees.

2.5.5 K-Nearest Neighbors (KNN)

K-nearest neighbors classifier is a supervised leaning method which uses clustering to classify. It can be summarized in three steps [23]:

1. Decide upon a number of k clusters and a distance metric.
2. Find the k number of nearest neighbors of a samples that is being classified.
3. Assign class labels by **majority voting**.

A new sample is assigned a class label based on the k closest samples by majority voting. Finding the right number of clusters to apply is key for a good performing KNN classifier.

2.6 Feature Importance

By computing feature importance, one can identify what impact a feature has on the model and, furthermore, filter out features that are redundant. This process result in dimension reduction and reduce the complexity of the dataset. There exist many types of feature importance methods, this thesis focus on permutation feature importance. Permutation feature importance simply consist of permuting the labels multiple times and compute the performance of the model by removing one feature at a time to find its importance.

Chapter 3

Materials

This section will elaborate more thoroughly on the structure of the dataset, and the composition of blocks (also discussed in chapter 4.3). The dataset is provided by Computational Radiology and Artificial Intelligence (CRAI), seated in Oslo University Hospital. The dataset contains 3873 patients with 1807 different measurements, which has been assembled from 2013 to 2020 from several hospitals in Norway.

3.1 Data collection

As mentioned, data from several hospitals were used for the collection of data. Different scanners for measurements of magnetic resonance imaging (MRI) images of the brain were used in different hospitals. These scanners are mostly from Siemens Healthineers such as Siemens Magnetom Prisma, Siemens Magnetom Avanto and Siemens Magnetom Skyra. Other scanners were provided from Philips, such as Philips Ingenia, Philips Intera and Philips Achieva. All scanners are considered to give the same results throughout this study.

Part of the data was also collected over the phone and through doctor appointments. Surveys were conducted over the phone, both for collecting new information and updating already existing information. Clinical data, such as blood tests and concentration of CSF beta-amyloid (betaA), was collected through doctor appointments.

3.2 Information on datasets

The number of patients used for the analysis is 172, and the number of original features is 113 from the main dataset. Some of these features were one-hot encoded, which will be elaborated in chapter 4.2. These features were divided into five blocks, which will be explained in this section. The features are divided and ordered after their relevance and similarity to each other.

3.2.1 Block A: background information

The first block consist of two features shown in table 3.1. These features have information about patients before any tests or other information are extracted.

Table 3.1: The table shows the first block used in the analysis. It contains categorical features of patients' background information.

Feature name	**Feature explanation**	**Data type**
recruit	Where the patient is recruited from, such as advertisement.	Categorical
subj_group	Which group a patient belongs to, such as control group or cognitive symptom group.	Categorical

3.2.2 Block B: cognitive and personal information

The second block contains information about patients' cognitive abilities through tests and personal information related to health and family relations. There are 16 features in this block, which is explained in table 3.2. The tests in this block are related to the ability to process and memorize words, which are used to determine if a patient is developing or having dementia or some cognitive impairment.

Table 3.2: The features in the second block. The block contains personal information and cognitive tests results.

Feature name	**Feature explanation**	**Data type**
gender	male or female	Categorical
smok	Smoker, no smoker or previous smoker	Categorical
cohab	cohabitation status	Categorical
marital	marital status	Categorical
edu_years	education years	Continuous
cowat_total	Controlled Oral Word Association Test	Continuous
age	The patient's age	Continuous
cerad_recall	Consortium to Establish a Registry for Alzheimer's Disease (CERAD) word list recall [37]	Continuous
cerad_recog	Consortium to Establish a Registry for Alzheimer's Disease (CERAD) word list recognition [37]	Continuous
cerad_learning	Consortium to Establish a Registry for Alzheimer's Disease (CERAD) word list memory [37]	Continuous
gds_score_comp	Geriatric depression scale (gds) categorized	Categorical

3.2.3 Block C: blood tests and cognitive tests

The third block, shown in table 3.3 contains cognitive tests used for cognitive assessment, such as the second block. In addition, this block contains blood test values and the Apolipoprotein E (APO-E) genotype mentioned in chapter 1.1. The tests in this block are standard tests used in the medical field to determine if a patient is developing or have dementia or a mild cognitive impairment. They are also used to determine other cognitive function impairment such as brain damage and the cognitive fitness for operating a vehicle.

Table 3.3: The third block, with features related to cognitive test results, blood test values and the Apolipoprotein E (APO-E) genotype.

Feature name	**Feature explanation**	**Data type**
bl_apoe	Apolipoprotein E alleles	Categorical
clock_score	Clock test used in the assessment of dementia	Continuous
bp_recum_sys	Systolic blood pressure when the patient is lying down	Continuous
bp_recum_dias	Diastolic blood pressure when the patient is lying down	Continuous
bp_1m_sys	Systolic blood pressure after 1 minute when the patient is sitting	Continuous
bp_1m_dias	Diastolic blood pressure after 1 minute when the patient is sitting	Continuous
bp_3m_sys	Systolic blood pressure after 3 minutes when the patient is sitting	Continuous
bp_3m_dias	Diastolic blood pressure after 3 minutes when the patient is sitting	Continuous
mor_ci	Mother with cognitive impairment	Categorical
mor_dem	Mother with dementia	Categorical
far_ci	Father with cognitive impairment	Categorical
far_dem	Father with dementia	Categorical
bmi	Body Mass index (BMI) value	Continuous
tmta_sec	Trail making test A (TMT-A) measured in seconds	Continuous
tmtb_sec	Trail making test B (TMT-B) measured in seconds	Continuous
mmse_total	Mini Mental Status Evaluation (MMSE) total score	Continuous
hyperchol	Hypercholesterolemia, also called high cholesterol	Continuous

3.2.4 Block D: lesion and white matter hyperintensity

The fourth block, shown in table 3.4, consist of lesion and white matter hyperintensity. This is a proxy for small vessel diseases in the brain. The "LES" features are divided into regions and layers. The four layers L1, L2, L3 and L4 are the division of the area between the cerebral cortex and the ventricles. The regions are cerebral lobes (FOPT), which are frontal lobe (F), occipital lobe (O), parietal lobe (P) and temporal lobe (T).

Table 3.4: The fourth block, with features of lesion (LES) and white matter hyperintensity (WMH).

Feature names	
Les_FPOT_rV	WMHo_rV
LesP1	LesP2
LesP3	LesP4
LesO1	LesO2
LesO3	LesO4
LesF1	LesF2
LesF3	LesF4
LesT1	LesT2
LesT3	LesT4
PSMD	

3.2.5 Block E: MR images of subcortical brain structures

The fifth and the last block contains volume measurements of subcortical brain structures with correction for intracranial volume and mean thickness of cortex areas from MR images. The features are shown in table 3.5 and table 3.6.

Table 3.5: The table shows thickness measurements in the given area of the cortex.

Feature names	
bankssts_thickness	caudalanteriorcingulate_thickness
caudalmiddlefrontal_thickness	cuneus_thickness
entorhinal_thickness	fusiform_thickness
inferiorparietal_thickness	inferiortemporal_thickness
isthmuscingulate_thickness	lateraloccipital_thickness
lateralorbitofrontal_thickness	lingual_thickness
medialorbitofrontal_thickness	middletemporal_thickness
parahippocampal_thickness	paracentral_thickness
parsopercularis_thickness	parsorbitalis_thickness
parstriangularis_thickness	pericalcarine_thickness
postcentral_thickness	posteriorcingulate_thickness
precentral_thickness	precentral_thickness
precuneus_thickness	rostralanteriorcingulate_thickness
rostralmiddlefrontal_thickness	superiorfrontal_thickness
superiorparietal_thickness	superiortemporal_thickness
supramarginal_thickness	frontalpole_thickness
temporalpole_thickness	transversetemporal_thickness
insula_thickness	MeanThickness_thickness

Table 3.6: The table shows volume measurements of subcortical brain structures with a correction for estimated intracranial volume.

Feature names	
LateralVentricle_rV	InfLatVent_rV
CerebellumWhiteMatter_rV	CerebellumCortex_rV
ThalamusProper_rV	Caudate_rV
Putamen_rV	Pallidum_rV
Hippocampus_rV	Amygdala_rV
Accumbensarea_rV	T1_Hippocampal_tail_rV
T1_subiculum_rV	T1_CA1_rV
T1_hippocampalfissure_rV	T1_presubiculum_rV
T1_parasubiculum_rV	T1_molecular_layer_HP_rV
T1_GCMLDG_rV	T1_CA3_rV
T1_CA4_rV	T1_fimbria_rV
T1_HATA_rV	T1_Whole_hippocampus_rV
Midbrain_rV	Medulla_rV
SCP_rV	Whole_brainstem_rV
AD_sig_surf_weighted	

Chapter 4

Methods

This chapter discusses the methodology and is divided into three sections; data preprocessing, data selection, and the primary analysis model; Sequential and Orthogonalised Partial Least Squares (SO-PLS). The summary of the method and the general workflow are shown in Figure 4.1.

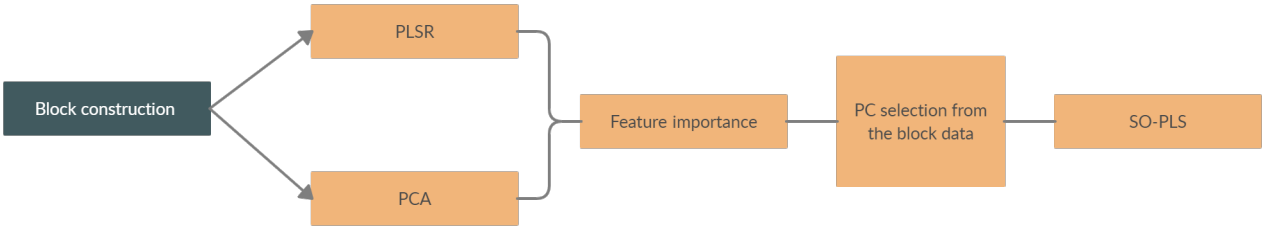


Figure 4.1: A summary of the workflow conducted, where PC is an abbreviation for principal component.

4.1 Software

This study used Python version 3.6.6 combined with Jupyter Notebook version 6.0.3 on the Anaconda platform with Hoggorm package [38] version 0.13.3 and Numpy [39] version 1.18.1. Other packages used are Scikit-learn [40] version 0.22.1, Pandas [41] version 1.0.0, HoggormPlot [38] version 0.13.2 and MLxtend [42] version 0.17.1.

4.2 Data preprocessing

The dataset was relatively prepared but had missing values. The preprocessing consisted of removing features with a high amount of missing values and standardizing the dataset. Features that were included in the analysis and were categorical, was transformed to one-hot numeric values.

4.3 Sequential and Orthogonalised Partial Least Squares (SO-PLS)

This section introduces Sequential and Orthogonalised Partial Least Squares (SO-PLS) regression, the main analytical regression method used in this study. SO-PLS is a supervised multi-block analytic model, which belongs to the area of component-based models such as PCA and PLS. There are some clear distinguishable differences between mentioned methods and SO-PLS, which is shown in table 4.1.

SO-PLS is an explorative technique that benefits from orthogonalization between multiple data blocks, which also manage to maintain the reliability of each data block and the overall variation to the dataset [19]. Uniquely for SO-PLS, data blocks with independent features are being added **sequentially** to the model to further explain the variance of the common response data [20]. Such a method gives a structured approach to complex datasets and may provide a better chance to accurately model the given phenomena [19].

Table 4.1: Overview of the component based methods PCA, PLS, SO-PLS and Multiple factor analysis (MFA). The table shows the difference between them in the number of predictor blocks or datasets, and the learning method (supervised and unsupervised).

	One data block	**Multiple data blocks**
Unsupervised learning	PCA	MFA
supervised learning	PLS	SO-PLS

The SO-PLS tries to divide the global datasets into blocks of variables that be-

long together in such way that it benefits the overall model by finding components that are optimal to each block. This is done by grouping features that explains the common part of the response variable or features that comes from the same sources into the same block. By doing this, we are taking into account the different complexities in different blocks which makes it possible to explain both the global variance and the block variance. Sequential modeling of the blocks can give us information about how much new information each block contributes to the understanding of the response we are trying to model.

4.3.1 Basic model of SO-PLS

The method will be discussed by using three X-blocks as input variables and Y as response variables, shown in formula 4.1. Where B, C and D denotes matrices containing parameters that are estimated, and E is a matrix containing errors or noise.

$$Y = X_1B + X_2C + X_3D + E \tag{4.1}$$

The method is based on first fitting X_1 to Y, and then on fitting the estimated residuals to X_2 after orthogonalization with respect to the extracted PLS components of X_1 for the first model. This is interpreted as that the only new information added to explain the variance in the response data from X_2 is the orthogonalized part of X_2. Which means already explained variance from X_1 will not been considered in the model [20].

The space spanned by X_1 and X_2 is the same as the space spanned by X_1 and X_2^{orth}, hence no loss of information in the process [20]. Where X_2^{orth} is X_2 orthogonalized with respect to X_1.

The SO-PLS method follows 5 steps [20];

1. Perform a simple PLS regression to fit X_1 to Y. Compute the X_1-scores, T_1, and the loadings for X_1 and Y, called P_1 and Q_1 respectively.
2. Compute the predictive model $T_1Q_1^T$ and the residuals, $E = Y - T_1Q_1^T$.
3. Orthogonalize X_2 with the scores and the principal components T_1 from step 1, and compute the orthogonalized X_2, X_2^{orth}.

4. Fit the residuals from step 2 to X_2^{orth} by using PLS regression, and compute new scores and loadings, T_2^{orth}, P_2^{orth} and Q_2^{orth}. Compute the predictive model $T_2^{orth}(Q_2^{orth})^T$.

5. Compute the final model, $\hat{Y} = T_1Q_1 + T_2^{orth}(Q_2^{orth})^T$.

The steps above apply for models with two blocks. Models with more than two blocks have to repeat the steps 3 to 5 with the residuals, scores and loadings from the previous estimated model.

4.3.2 Selection of components for each block

The SO-PLS procedure require selection of principal components of each block before handling the final model. The selection of components of each block is done sequentially by choosing the amount of components that gives the lowest root mean squares error of prediction (RMSEP). The combination of component is shown through a RMSEP-plot based on a cross validation (CV), also known as Maage plot. An example of such a plot is shown in Figure 4.2.

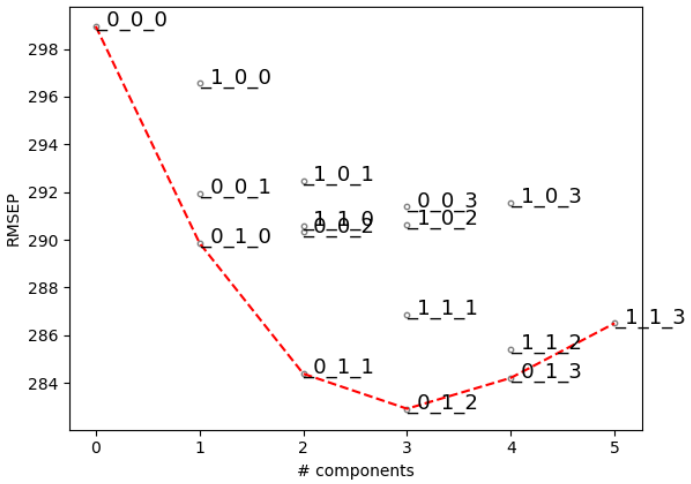


Figure 4.2: RMSEP plot showing 3 blocks with different combination of principal components, where the numbering order represent principal components from each block accordingly.

For choosing components for a new block, the components from previous blocks are kept constant. In Figure 4.2, the components from the first and second block are

one. With the assistance of the RMSEP plot, we decide the number of components to use from the third block; in Figure 4.2, we choose two components. The final model is then one component from the first and second block and two components from the third block. The procedure is explained in the steps below;

1. Perform PLS on the first block, and construct and visualize the RMSEP for each components. Choose the amount of components that gives the lowest RMSEP value.

2. Perform SO-PLS with the first and second block, and construct a RMSEP plot. Keep the amount of components from the first block the same and choose the amount of components for the second that gives the lowest RMSEP value.

3. Repeat step 2 for the next blocks potentially.

4.4 Data selection

The methods for selection of the final dataset will be elaborated in this section, for explanation of the dataset see chapter 3.

4.4.1 Organising features in blocks

The organisation of feature in different blocks was based mainly on source or origin. Some example of source or origin for this dataset is, among others, MRI, blood test and cognitive tests, and methodology is questions through phone calls and cognitive tests.

To verify that the blocks did not have a high correlation to each other, the R_V-coefficients were calculated. The R_V-coefficient measures the closeness of two vectors that are each represented in a matrix. By examine the R_V-coefficients, further changes to the blocks can be made.

The order of the blocks in the model is also important for the analysis. The block order was decided after each block's data accessibility and origin. The blocks was sorted in a chronological order, where the first block consist of background information, second block of cognitive and personal information, third block of

blood tests and cognitive tests, fourth block of lesions in the brain, and fifth block of volume measurements from subcortical brain structures.

4.4.2 Feature selection and data cleaning

The number of patients was determined after missing values in the selected features. All patients with missing values, such as NaN (not a number) and the values showing -999, were removed from the dataset. This decision was made after finding the right amount of features in the given block. Features that had a significant percentage of missing values was calculated to decide the importance of the feature empirically.

PCA loadings were used for each block and for the whole dataset to find features that explained most variance. These features were also chosen empirically and included in the model. PCA was performed several times with different blocks and features in an explorative manner.

Feature importance permutation from the MLxtend package was used on each block separately to decide some features importance to the analysis. The response value was simplified to a binary value where values between 600 and 800 were removed. In other words, feature importance permutation was performed on a binary classification problem to decide which features to remove from the model.

Chapter 5

Results

This chapter will review the general outcome of the analysis completed, and a more detailed review will be covered in chapter 6. Section 5.1 will showcase the assessment of features and analysis methods used for the selection of features, and the reasoning of the placement of features in different blocks. The process of establishing the final model will be reviewed in section 5.2. Model performance and the outcome of the model will be reviewed in section 5.3.

5.1 Data preparation and pre-analysis

The selection of features is based on some analytical methods and different sources. The methods used were PCA, PLS, and feature importance on each block individually to estimate the explained variance in the block and evaluate if the features in each block give a sufficient value of the response's explained variance. In this section, the results from PLS, PCA and feature importance will be reviewed. The next subsections present the analysis done on each block separately.

Partial least square regression and principal component analysis was conducted before the main analysis of SO-PLS. The results are shown individually in the sections below. Visualization of each methods was conducted such as correlation loadings plots, loadings plots and plots of explained variance of the given block and the response.

Feature importance permutation was also performed for each block. The results of

these permutation was tested against PLS to find the combination of features that gave the highest validated explained variance. Due to too many plots constructed from PLS, PCA and feature importance permutation, many of these will not be visualized. A summary of PCA and PLS are showcased in table 5.1 and table 5.2. From the tables mentioned, 3 components was set as the maximum number of components.

Table 5.1: The table shows a summary of the blocks explained variance extracted from the PLSR. These values shows the cumulative explained variance with given number of blocks for both calibrated and validated data. X represent the block data, while Y represent the response vector. The results were set to maximal 3 components.

Block	**Calibrated exp. var. of X**	**Validated exp. var. of X**	**Calibrated exp. var. of Y**	**Validated exp. var. of Y**	**#components**
A	45.92%.	45.05%	7.28%	5.73%	3
B	31.45%	30.16%	18.42%	14.69%	3
C	30.82%	29.95%	17.60%	13.38%	3
D	72.72%	69.41%	14.39%	4.20%	3
E	64.85%	64.24%	25.55&	16.83%	3

Table 5.2: The table shows a summary of the blocks explained variance extracted from the PCA. These values shows the cumulative explained variance with given number of blocks for both calibrated and validated data, X. The results were set to maximal 3 components.

Block	**Calibrated exp. var. of X**	**Validated exp. var. of X**	**#components**
A	51.79%.	50.68%	3
B	41.65%	40.48%	3
C	50.93%	49.78%	3
D	74.51%	77.61%	3
E	62.58%	62.13%	3

5.1.1 PLS

PLS was conducted on each block, because of the sheer number of plots, this section will showcase the result of PLSR of block B. The figures 5.1, 5.2 and 5.3 shows the correlation loadings plot, explained variance and prediction plot respectively. PLS was done to detect the explained variance, thus the information in each block separately. It was also done for outlier detection, both for features and values in each block. These outliers would later on examined to find drift, or mistakes, in the data or to find important information.

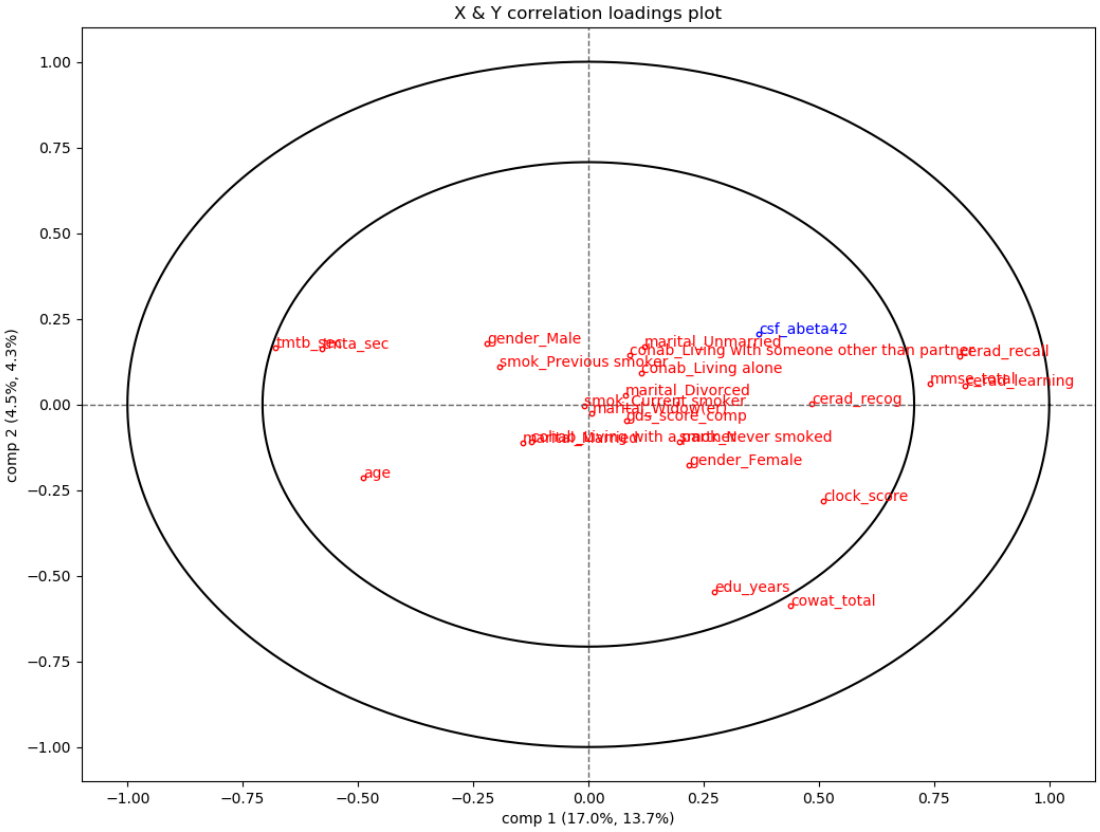
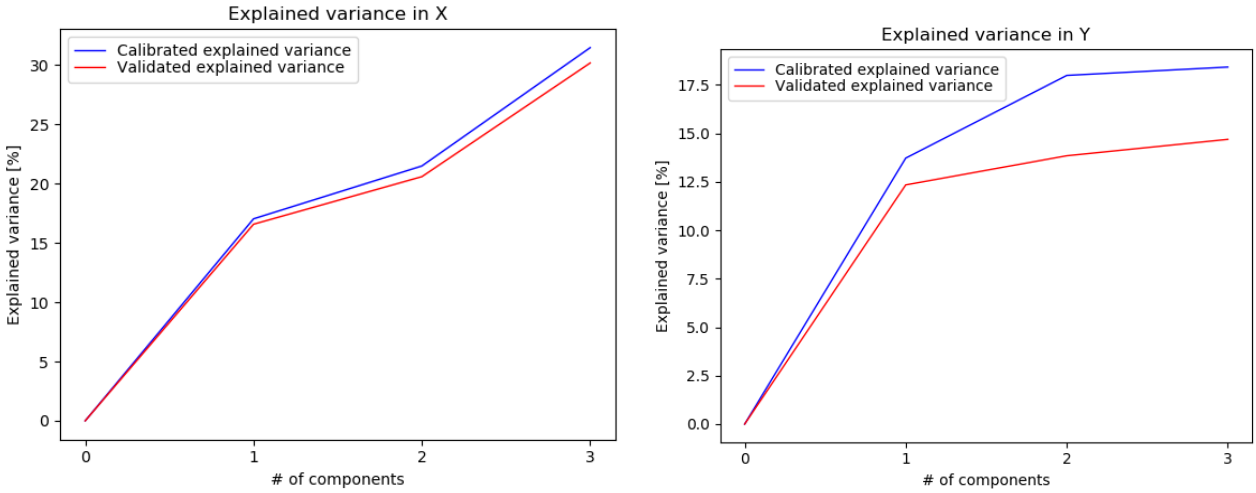


Figure 5.1: Correlation loadings plot of the response in blue and categorized features from block B in red. The horizontal axis shows the first principal component and the vertical axis shows the second principal component. The sum of component 1 and component 2 explains $17\% + 4.5\%$ of the variance in X, and explains $13.7\% + 4.3\%$ of the variance in the response (csf_abeta42). The inner circle in the plot is the 50% threshold for the explained variance. Features inside of this circle has an explained variance of 50% or lower. The outer circle represents 100% explained variance. Some features, such as cowat_total and cerad_recall, are between 50% and 100% explained variance.



(a) **(b)**

Figure 5.2: Explained variance plot of block B, (a), and of the response, (b). The horizontal axis represent number of principal components computed by a PLS regression, and the vertical axis represent percentage of explained variance, both calibrated and validated.

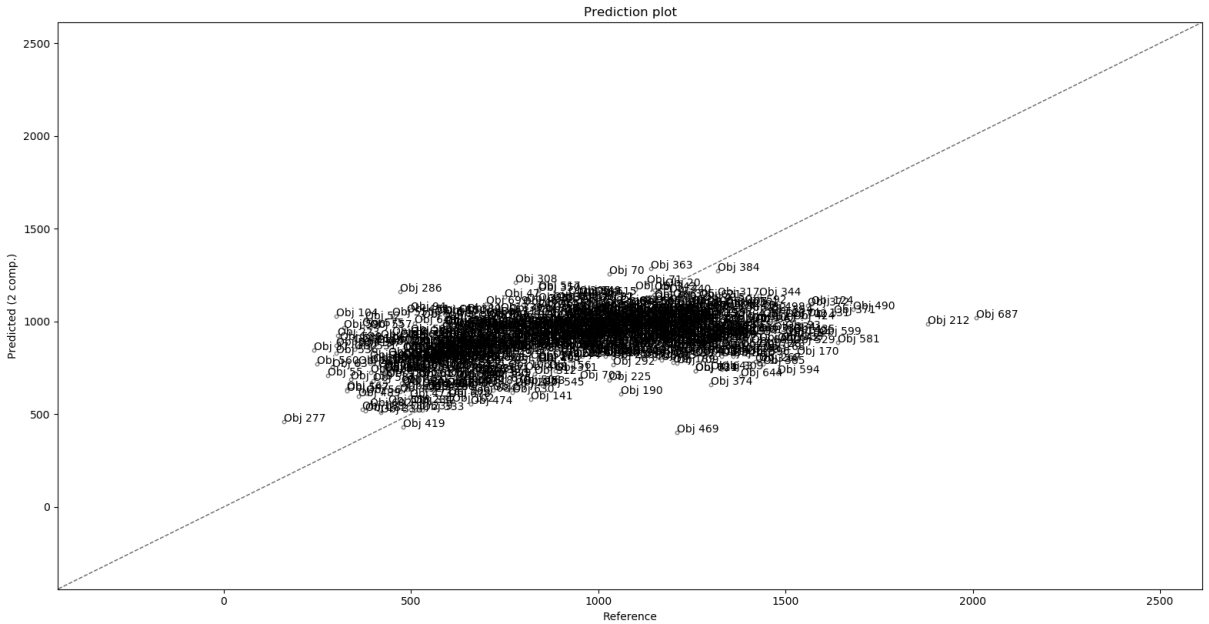


Figure 5.3: Prediction plot of block B. The horizontal axis shows the real values, while the vertical axis shows the predicted values based on two components.

5.1.2 PCA

Just as PLS, PCA was conducted on each block, this section will showcase the result of PCA of block E. The figures 5.4, 5.5 and 5.6 shows the correlation loadings plot, the explained variance and the scores plot respectively. PCA was applied to examine and detect the information and the variance in each block. Such as with PLS, PCA was conducted to find outliers for better understanding of important aspect of the data, and to remove faulty data.

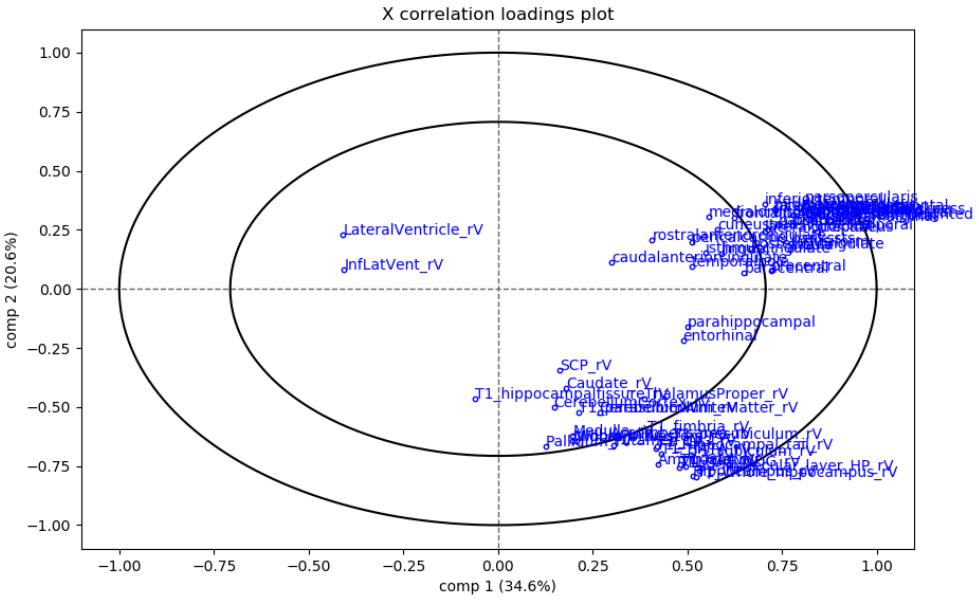


Figure 5.4: Correlation loadings plot of the categorized features from block E. The horizontal axis shows the first principal component and the vertical axis shows the second principal component. Most of the features are clustered together, which makes them correlated to each other. The figure shows two clusters with correlated features.

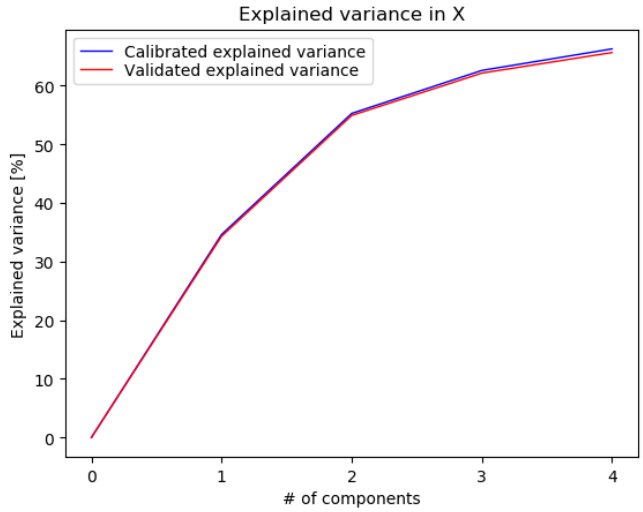


Figure 5.5: Explained variance plot of block E. The horizontal axis represent number of principal components constructed by PCA, and the vertical axis represent percentage of explained variance.

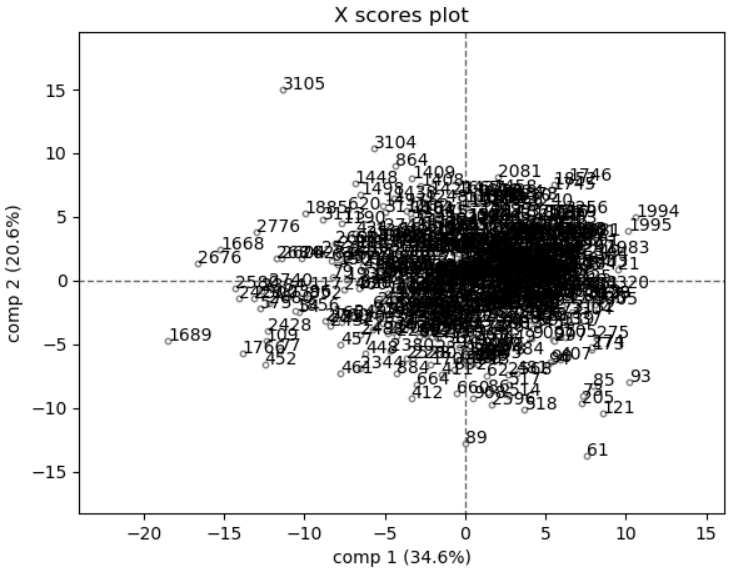


Figure 5.6: Scores plot of block E. The first component is the horizontal axis and the second component is the vertical axis. Component 1 and component 2 explains 34.6% and 20.6% of the total variance of block E respectively, that is more than half of the variance is explained by two principal components.

5.1.3 Feature importance permutation

Feature importance permutation was the last part of the preprocessing and pre-analysis process. Each block was tested against 5 machine learning models; support vector machine, logistic regression, decision tree, random forest and K-nearest neighbor (see section 2.5). The response value was transformed to a binary response value, where values between 600 and 800 of the response value (beta-amyloid) was removed. The feature importance method was then conducted for each block.

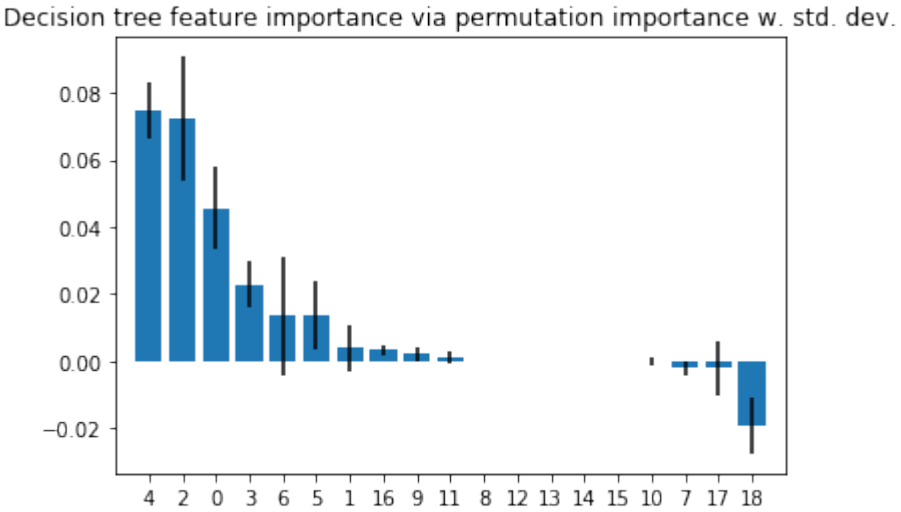


Figure 5.7: Decision tree feature importance permutation of block C with categorized features. The features are represented by index number shown in the horizontal axis.

Figure 5.7 shows the outcome of one feature importance permutation for block C. Models that gave an accuracy score above 70% were tested separately in PLS to find which combination of feature that gave the highest validated explained variance. Features that had a score between 0.001 and -0.001 in the feature permutation was removed from the dataset. The final result is shown in table 5.3.

Table 5.3: The table shows the result of the feature importance permutation that was conducted as the final part of pre-analysis. The model shown are the model that computed the highest validated explained variance with PLS.

Block	**model**	**Validated exp. var. of Y**	**#components**
A	SVC	5.78%	3
B	Decision tree	15.68%	3
C	Decision tree	12.48%	3
D	Random forest	6.43%	1
E	Logistic regression	16.43%	3

5.1.4 Block selection

The reason for the block order and feature selected is based on several factors. First and foremost, the block order was determined after the complexity and accessibility of the data mining process; data from block A is easier to gather and access than data from Block E. The first measurements medical experts perform on a patient are gathered in the first blocks, as measurements get more complicated and difficult to provide. The feature of these measurement gets moved to a later block. By the mentioned reasoning, the separation of data blocks was done after where in the process the measurement was obtained, hence the origin or the source of the measurement.

Block features internal relevance to each other was also an important factor where features were similar and difficult to separate. There was a trial and error process, in some degree, where blocks of some similarities were altered several times to maximize explained variance tested with PLS and PCA. The features were also selected in consultation with medical experts with the domain knowledge, that is the providers of the dataset.

5.2 Final SO-PLS model

The creation of the final model is conducted through principal component selection for each block according to the method of SO-PLS, explained in chapter 4. The final model with the principal components is **[1, 2, 2, 3, 0]**, where the values represent principal components of the each block in alphabetically order. The component selected are shown in RMSEP plots below. The same patients across blocks are used for SO-PLS, thus the number of patients used for SO-PLS is 172.

The selection of the number of components is dependent on the previous selected components from the previous block, e.g. the component selected from block A are fixed when selecting components from block B. The figures below shows the RMSEP plot of each block and the final selection of principal components.

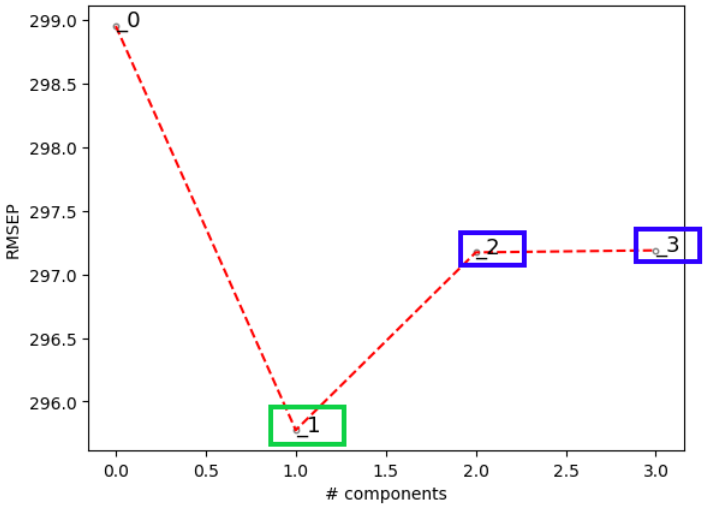


Figure 5.8: RMSEP plot of block A, where relevant number of components are marked in blue and number of components selected are marked in green. The number of components selected are 1.

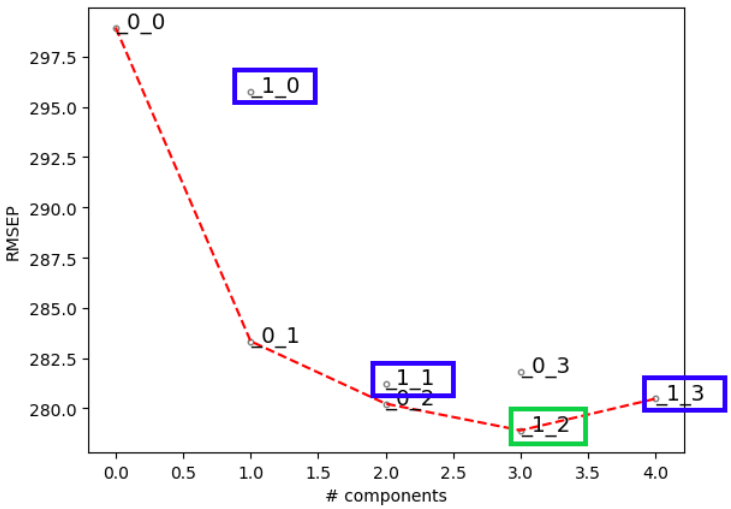


Figure 5.9: RMSEP plot of block B combined with block A, where relevant number of components are marked in blue and number of components selected are marked in green. The number of components with the lowest RMSEP are selected, from block B that is 2 components.

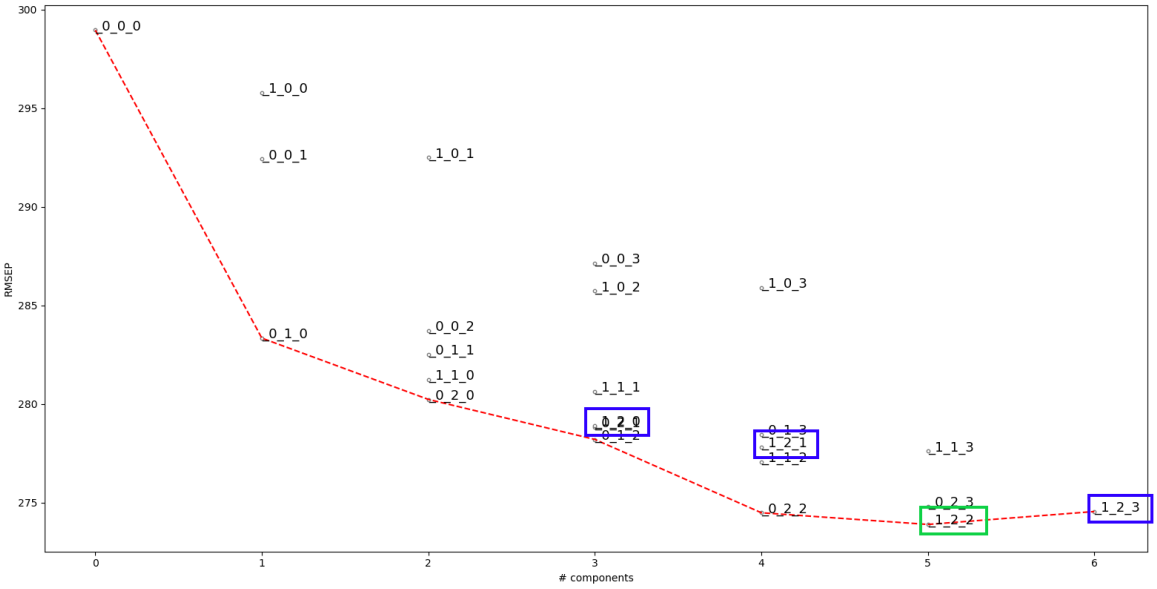


Figure 5.10: RMSEP plot of block C combined with block A and B, where relevant number of components are marked in blue and number of components selected are marked in green. The number of components with the lowest RMSEP are selected, from block C that is 2 components.

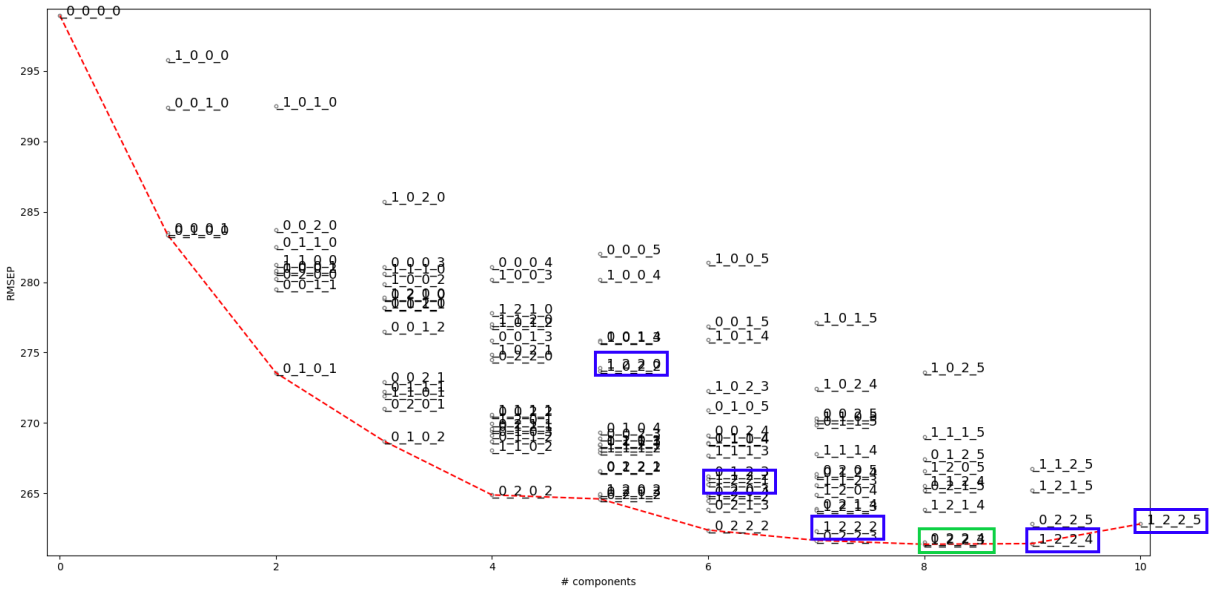


Figure 5.11: RMSEP plot of block D combined with block A, B and C, where relevant number of components are marked in blue and number of components selected are marked in green. The number of component selected from block D are 3.

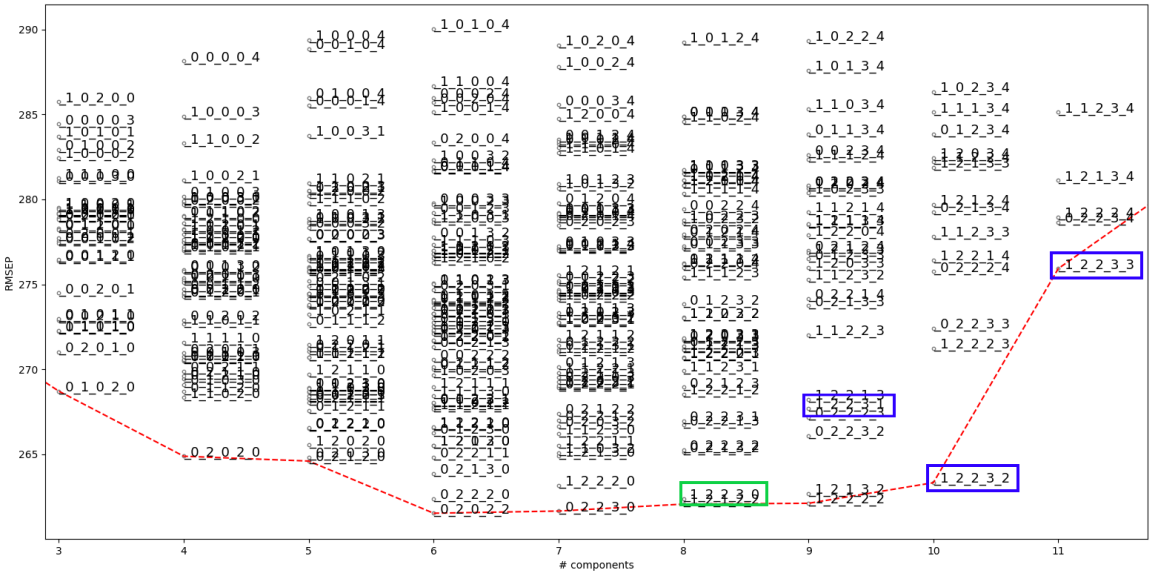


Figure 5.12: RMSEP plot of block E combined with block A, B, C and D, where relevant number of components are marked in blue and number of components selected are marked in green. The number of components with the lowest RMSEP are selected, from block E that is 0 components. To avoid overfitting the final model, selecting zero components is the preferred choice.

5.3 Model performance

The final model's performance and outcome are shown in Figure 5.13, 5.14 and 5.15. The validated explained variance of the response was **22.89** percent, the calibrated is **36.54** percent. Each block's contribution to the explained variance are shown in table 5.4.

Table 5.4: The table shows the explained variance contribution from each block.

Block	**Calibrated exp. var.**	**Validated exp. var**
A	5.76 %.	2.52 %
B	14.95 %	10.22 %
C	7.47 %	2.90 %
D	8.36 %	7.25 %
E	0 %	0 %
Total	36.54 %	22.89 %

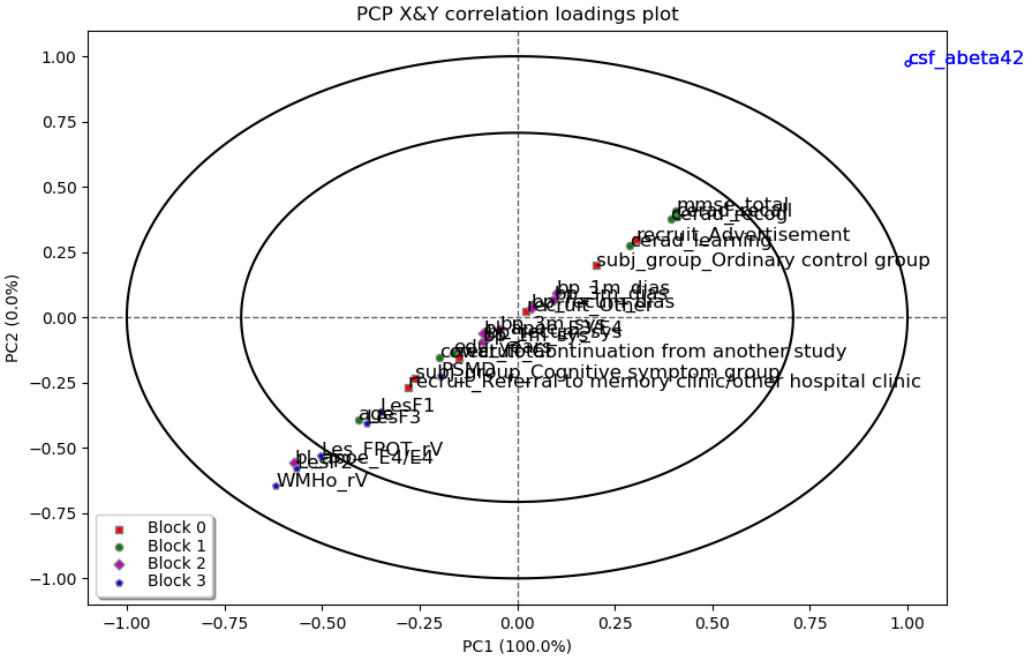


Figure 5.13: The figure shows a correlation loadings plot of the data features and response. The features from same block are categorized with the same color. The axis represent the first and second principal component, and their contribution in parenthesis. Block 0 represent the first block added to the model, block 1 represent the second block and so on. The features closest to the response, csf_abeta40, are the features that are highest correlated with the response. Features that are outside the inner circle have an explained variance above 50%. Such features are WHMo_rV and bl_apoe E4/E4 shown in the lower left panel of the plot.

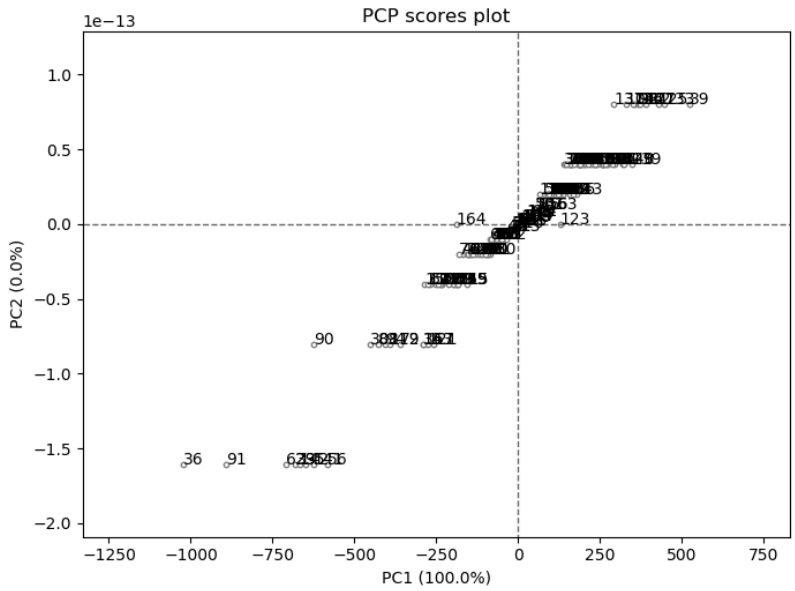


Figure 5.14: The figure shows a scores plot of the data values with the first and second principal component as the axis.

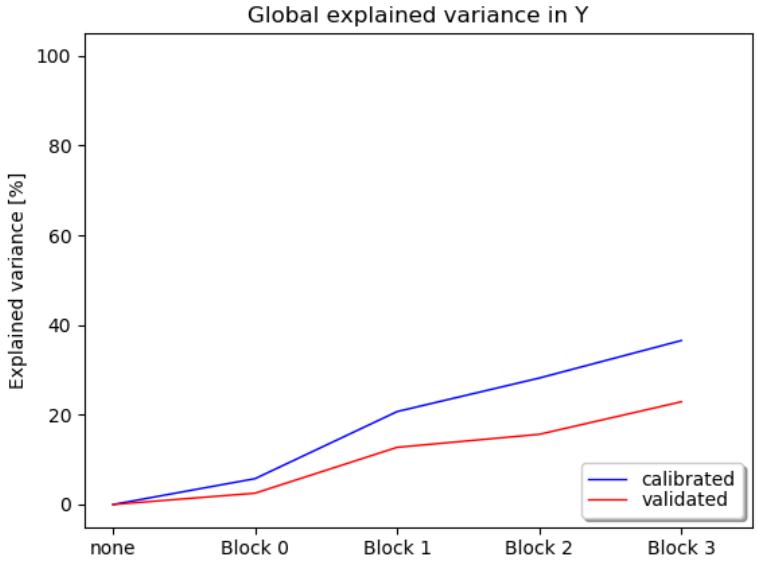


Figure 5.15: The plot shows the calibrated and the validated explained variance accumulated after each block is added to the model. Block 0 represent the first block, block A and so on. The plot shows that there is a high amount of explained variance that the model do not explain.

Chapter 6

Discussion

6.1 Dataset

Alzheimer's is a progressive disease where symptoms worsen over several years. These symptoms are difficult to detect in the early stages, making the data collection and data mining process difficult for medical experts. This dataset is no exception from that, as for now, there exist some dominant and clear risk factors, and the dataset utilized in this thesis found weak links between known risk factors and AD. This section discusses the data used and its weaknesses.

6.1.1 Features

There are 3873 patients with 1807 different measurements in the dataset; only a few of these features were used. There may be important features not used for the final SO-PLS model, which is a cause of limited knowledge in the existing dataset and the medical field. Finding important features from the original dataset is difficult because of the sheer number of missing values of various features, which do not apply to the methods used. The number of patients also significantly drops when removing patients with a large amount of missing values.

By including more features to the methods that do not necessary correlate with AD significantly, the methods may have a bigger change to find hidden patterns that contributes to the assessment of AD. In contrast, including more features may also lead to overfitting the model, since there is a chance that some features

will introduce noise in the datasets and not contribute to the improvement of the model. Adding more features should still have been conducted to find new risk factors of AD since there is insufficient amount of information from features in the existing model. The information that is absent in the model may exist in other features or measurements.

6.1.2 Response variable

The final model is an univariate regression problem, with the response variable measuring the accumulation of CSF beta-amyloid in the spinal fluid. The measurement of tau mentioned in chapter 1.1 was also part of the original dataset. Tau could also be used for the analysis as a second response variable and tested with various models and analysis. The reason tau was not used is because of the high correlation with beta-amyloid. There was no new information added when tau was included, and the model was not improved. Tau may still be an important variable due to its importance in deciding the degree of a patient's dementia.

The dataset consists of patients that have been tested for AD through several stages. These patients that went through these assessments have a higher probability of having AD than the rest of the population pool. There is, therefore, a lack of patients that shows little to none sign of AD. Adding data of healthy people to the datasets and finding patterns that are unique to AD patients may be easier, making it easier to differentiate between healthy and AD patients. Furthermore, the features of importance would be more detectable and easier to identify.

6.1.3 Block selections

The order of blocks and the grouping of features in the respective blocks are important when designing a good model. The selection of blocks and features was based on dominant features from previous studies, their relation to each other, and the block's order. The block selection and order were explorative through regular PLS and PCA models and highly influenced by the features' origin. The blocks' division and order were designed by medical experts with strong domain knowledge in the datasets provided. Furthermore, the block division could have been based on a stronger knowledge in medical practices and, more specifically, in Alzheimer's disease and the main datasets with the SO-PLS method in consideration.

A clearer view of the method's advantages and applicability with insight to the

datasets' internal differences from a top-down approach, rather than a bottom-up approach, where features got selected and placed in their respective block may cause a better separation of the block. Instead of selecting features first and placing them in their respective block, one could define each block's property and find features that fit a block's description. This would cause a better block separation and make it easier to search for specific features for a block.

6.2 The final model and performance

The outcomes from PLS showed validated explained variances of the response to be up to 20% for each block, which individually is not very robust models. The expectation for further analysis using SO-PLS was that all blocks together might explain more of the variance of the response than each block individually. This was partly the case since each block, excluding block E, contributed to a higher explained variance. however, this was not as much as expected. The reason for the low explained variance is overlapping information of the blocks. If each block contained unique information of the response, the total explained variance of the response for the SO-PLS model would have been much higher.

Block E was not used for the SO-PLS method despite having the highest percentage of explained variance from PLS. This shows that block E's explained variance is explained by the other blocks in the dataset, thus making block E redundant. Block E still holds information about the response, but since the block's order was preset before the analysis, the block became unnecessary.

Such as explained earlier, the final model may also suffer from insufficient data and poor block separation. The model may experience the "garbage in, garbage out" concept, where the data consist of a large amount of noise and useless information and therefore gives a poor model. Since both the data and its separation is fundamental for a better model, this may be the case in this model. There is also a possibility that the dataset does not contain relevant features that explain the variance in the response that the information available is not enough to explain the variance in the response. There may be a need for new features that have not been included at all.

Another drawback with the final model is the amount of patients. Since SO-PLS requires the same patients to exist in all blocks, patients not included in all blocks are removed from their respective blocks. The final model ends up with 172 patients across all blocks, which may be an insufficient amount for the overall

model. More data may give more patients to the final model, or more information on already existing patients to reduce missing values may also contribute to a better model.

The features WMHo_rV, bl_apoe E4/E4, LesP2 and Les_FPOT_rV had an explained variance above 50% in the final model (see Figure 5.13). The high explained variance from these features indicates a negative correlation between the response and the features mentioned above, indicating that high beta-amyloid (low CSF betaA) gives high values for the given features. As mentioned in Chapter 1.1, high values of beta-amyloid (betaA) gives low values of CSF betaA (csf_abeta42) in the spinal fluid.

The feature bl_apoe E4/E4 is the ApoE-ϵ4 allele mentioned in Chapter 1.1. This allele frequency is much higher for patients with AD, which agrees with Figure 5.13. LesP2 is the lesion in the second layer of the parietal lobe. LesP2 also show a negative correlation with the response value, which is shown in Figure 5.13. The parietal lobe has been proven to be involved in the early stages of AD and mild cognitive impairment [43], and are, therefore, another important feature in this model. The same reasoning also applies for WMHo_rV [44] and Les_FPOT_rV [45], where Les_FPOT_rV is an aggregated variable of lesion FPOT with corrections for intracranial volume and WMHo_rV is areas of low-intensity white matter calculated from histograms from MRI.

In consideration of the correlation loadings plot shown in Figure 5.13, the first component (PC1) explains 100% of the total variance making the second component (PC2) equal to 0% explained variance. The figure shows by equal distribution of features along both axis while, in theory, it should be a wider span across PC1. By closer examination of the scores plot in Figure 5.14, there is a similar distribution. The difference is shown in the magnitude of the axis in the scores plot; there is a much larger span in PC1 than in PC2.

High age has been shown to give a higher probability of developing AD. The feature age gave an explained variance under 50% in the loadings correlation plot shown in Figure 5.13. The reason for this may be the distribution of the patients, which is shown in Figure 6.1. There exist no patients under the age of 40, which gives age a lower effect on the model. Since the main goal is the find risk factors and features that contribute to AD, the early effects of patients are eliminated since no patients under the age of 40 are included. To find early effects of AD, one may consider adding a younger age group to the dataset.

By examine age and csf_abeta42 more thoroughly, it shows a decline in csf_abeta42

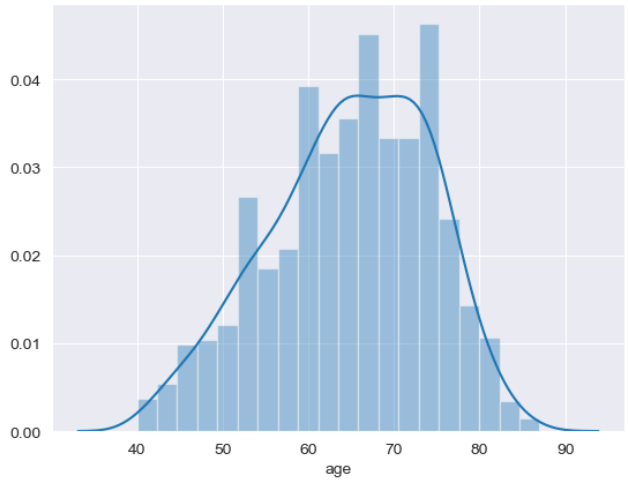


Figure 6.1: Distribution plot of the feature age from 1518 patients of the general dataset.

as the patients' age gets higher. A simple illustration of this trend is shown in Figure 6.2. The age appears to be more important than shown in the final model. By including more data, age shows a more significant difference across age groups. This may be the cause of insufficient data since the SO-PLS used 172 patients across all blocks. More patients across a wider age range may show the importance of age and other features.

SO-PLS is a linear model, which means that features that behave in a non-linear matter may not affect the final model. One may introduce features' non-linearity by producing squares of features, either with itself or pairwise with other features. Figure 6.2 shows some tendency to a non-linear variable, making the feature's importance to the model lower than its true contribution. Because of the time limit, these changes were not implemented.

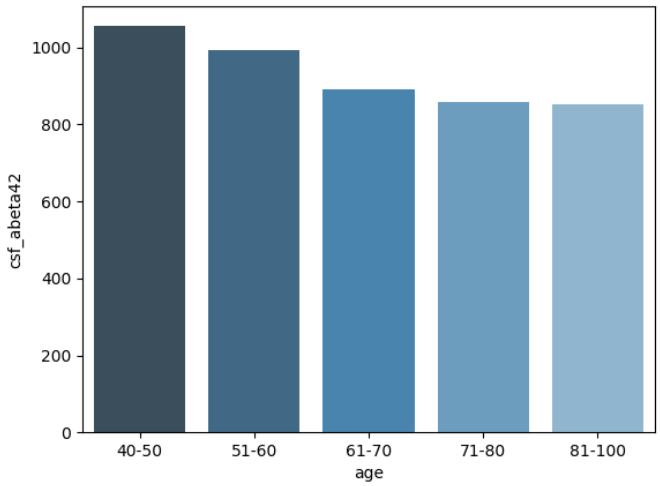


Figure 6.2: Bar plot of the feature age with csf_abeta42 from 1518 patients of the general dataset. The plot shows the mean of csf_abeta42 for each age group.

6.3 The aim of this thesis

The aim of this thesis was to apply PCA, PLS, and SO-PLS to determine which factors or variables improve the assessment of a patient with AD. There were few known modifiable risk factors of AD, and this thesis amplifies this known state. The validated explained variance of the response for SO-PLS is 22.89%, which is not sufficient for determining significant features and risk factors. However, there are still some noticeable tendencies.

The goal was to determine features and risk factors for AD; this does not necessarily indicate measuring the response, csf_abeta42, for concluding if a patient has AD. For determining if patients have AD, some clinical criteria must be fulfilled. As explained in Chapter 1.1, these criteria are often carried out and checked by medical experts in the field, and from those assessments made a decision if a patient has AD or other cognitive diseases. The current response value may, therefore, not be a precise representation for deciding if a patient has AD. Additional measurements, such as MRI scans may be an alternative approach to get a more exact response value.

6.4 Further work

SO-PLS is not a widely applied method in medical science and has the potential for further work in the medical field. One of the benefits of SO-PLS in the medical field is utilizing the distribution in the scores plot and correlation loadings plot to get a better insight into the patient pool. By examining the scores plot, it is possible to see which patients are similar and which one differs. For doctors to utilize this means to diagnose and tailor the patients' treatment across hospitals and treatment centers. It is a tool that can assist doctors when, in doubt, by comparing similar patients in the scores plot. This also may be beneficial when other dementia patients are included in the dataset to make it easier to separate similar diseases. There is also a possibility to apply the scores plot alongside the correlation loadings plot to examine which features affect the patients the most. Finding important features and how those features affect the variance for a particular patient will give a better insight into the patients.

Alzheimer's disease is yet to be fully explained, and thus are difficult to predict and to analyze. As big data and machine learning become more applicable in various fields, data sharing, data mining, and cooperation become necessary to construct sufficient datasets. With the help of progresses in the medical field, such initiatives will change the outlook of Alzheimer's disease. Initiatives such as Alzheimer's Disease Data Initiative (ADDI) will accelerate the progress towards a breakthrough in Alzheimer's disease.

For further analysis, it should be taken into consideration that some assessments may be outdated and not beneficial to use. Such features may be memory assessment that uses analog clocks, such as clock_score used in block C. The mentioned feature may be difficult for a younger population as analog clocks become unused. More importantly, is to use new features and assessments that add new information to the data and that do not include redundant or similar assessments. Digital tools and assessments, such as puzzles and other mind games, may contribute to new hidden information.

Chapter 7

Conclusion

The overall goal was to find important risk factors and features to get a deeper understanding of why some develop AD. Mainly the methods PCA, PLS and SO-PLS in conjunction with feature importance, was applied to detect important features. The final model, SO-PLS, gave an explained variance of the response of 22.89%, which means that 22.89% of variance is explained by the given model. From the model's outcome, the features WMHo_rV, bl_apoe E4/E4, LesP2 and Les_FPOT_rV had a negative correlation with the response, csf_abeta42, thus positive correlated with beta-amyloid in the brain. The mentioned features had an explained variance between 50% and 100%, and are therefore considered important features.

The methodology, SO-PLS, did not reach its full potential in this thesis because of poor performance and the lack of sufficient data and good block separation. The model still has the potential to be tested with AD data, and in general, with clinical and medical data. The SO-PLS model showed some promising features, but the low explained variance makes these results highly questionable and concludes that no decisive factors can be given to what causes AD.

A larger dataset may contribute to a better final model for further work and should be considered when utilizing the given dataset. Several other features that were not utilized in the blocks may explain other components of AD and should be considered for further work. The separation of data into data blocks should be fully understandable and reasonable, and other reasoning used in this thesis should be explored.

Bibliography

[1] W. H. Organization, *Global action plan on the public health response to dementia 2017 - 2025*. 2017, ISBN: 978-92-4-151348-7.

[2] J. P. NV. (). Dementia prevention, [Online]. Available: https://www.dementia.com/prevention.html#refrences. (accessed: 06.06.2020).

[3] N. I. o. H. National Institute on Aging, *Comparison of a healthy brain and a brain with severe alzheimer's disease*, 2016.

[4] K. Engedal. (). Alzheimers sykdom, [Online]. Available: https://sml.snl.no/Alzheimers_sykdom. (accessed: 28.05.2020).

[5] B. C. Riedel, P. M. Thompson and R. D. Brinton, 'Age, apoe and sex: Triad of risk of alzheimer's disease', *The Journal of Steroid Biochemistry and Molecular Biology*, vol. 160, pp. 134–147, 2016, SI:Steroids & Nervous System, ISSN: 0960-0760. DOI: https://doi.org/10.1016/j.jsbmb.2016.03.012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960076016300589.

[6] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. J. Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub and C. H. Phelps, 'The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease.', *Alzheimers Dement*, vol. 7(3), pp. 263–269, 2011. DOI: 10.1016/j.jalz.2011.03.005. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21514250/.

[7] Helsedirektoratet. (). Demens - nasjonal faglig retningslinje, [Online]. Available: https://www.helsedirektoratet.no/retningslinjer/demens. (accessed: 29.06.2020).

[8] L. CC, L. CC, K. T and B. G. Xu H, 'Apolipoprotein e and alzheimer disease: Risk, mechanisms and therapy', *Nature reviews. Neurology*, vol. 9(2), pp. 106–118, 2013. DOI: 10.1038/nrneurol.2012.263. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3726719/#R10.

[9] T. Burt, B. Agan, V. Marconi, W. He, H. Kulkarni, J. Mold, M. Cavrois, Y. Huang, R. Mahley, M. Dolan, J. McCune and S. Ahuja, 'Apolipoprotein (apo) e4 enhances hiv-1 cell entry in vitro, and the apoe $\epsilon 4/\epsilon 4$ genotype accelerates hiv disease progression', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 8718–8723, 2008. DOI: 10.1073/pnas.0803526105. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2438419/.

[10] C.-L. Kuo, L. C. Pilling, J. L. Atkins, J. A. H. Masoli, J. Delgado, G. A. Kuchel and D. Melzer, 'Apoe e4 genotype predicts severe covid-19 in the uk biobank community cohort', *The Journals of Gerontology: Series A*, 2020. DOI: https://doi.org/10.1093/gerona/glaa131. [Online]. Available: https://academic.oup.com/biomedgerontology/advance-article/doi/10.1093/gerona/glaa131/5843454.

[11] H. MN, S. L, J. WJ, M. TA and K. L., 'The role of apoe $\epsilon 4$ in modulating effects of other risk factors for cognitive decline in elderly persons', *JAMA*, vol. 282(1), pp. 40–46, 1999. DOI: 10.1001/jama.282.1.40. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10404910/.

[12] P. R, R. BL and L. LJ., 'Type 2 diabetes, apoe gene, and the risk for dementia and related pathologies: The honolulu-asia aging study.', *Diabetes*, vol. 51(4), pp. 1256–1262, 2002. DOI: 10.2337/diabetes.51.4.1256. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11916953/.

[13] M. Malek-Ahmadi, K. Davis and C. Belden, 'Validation and diagnostic accuracy of the alzheimer's questionnaire', *Age Ageing*, vol. 41(3), pp. 396–399, 2012. DOI: 10.1093/ageing/afs008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22367356/.

[14] A. o. h. Nasjonal kompetansetjeneste. (). Skalaer og tester, [Online]. Available: https://www.aldringoghelse.no/skalaer-og-tester/. (accessed: 08.12.2020).

[15] T. Tapiola, I. Alafuzoff, S. K. Herukka, L. Parkkinen, P. Hartikainen, H. Soininen and T. Pirttilä, 'Cerebrospinal fluid beta-amyloid 42 and tau proteins as biomarkers of alzheimer-type pathologic changes in the brain', *Archives of neurology*, vol. 66(3), pp. 382–389, 2009. DOI: https://doi.org/10.1001/archneurol.2008.596. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19273758/.

[16] N. Hill and J. Mogle, 'Alzheimer's disease risk factors as mediators of subjective memory impairment and objective memory decline: Protocol for a construct-level replication analysis', *BMC Geriatr.*, vol. 18(1), p. 260, 2018. DOI: 10.1186/s12877-018-0954-5. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30373526/.

[17] C. Qiu, M. Kivipelto and E. von Strauss, 'Epidemiology of alzheimer's disease: Occurrence, determinants, and strategies toward intervention', *Dialogues in clinical neuroscience*, vol. 11(2), pp. 111–128, 2009. DOI: 10.1093/ageing/afs008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181909/#.

[18] S. Wold, K. Esbensen and P. Geladi, 'Principal component analysis', *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, ISSN: 0169-7439. DOI: https://doi.org/10.1016/0169-7439(87)80084-9. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0169743987800849.

[19] J. Niimi, O. Tomic, T. Næs, D. W. Jeffery, S. E. Bastian and P. K. Boss, 'Application of sequential and orthogonalised-partial least squares (so-pls) regression to predict sensory properties of cabernet sauvignon wines from grape chemical composition', *Food Chemistry*, vol. 256, pp. 195–202, 2018, ISSN: 0308-8146. DOI: https://doi.org/10.1016/j.foodchem.2018.02.120. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308814618303613.

[20] B.-H. M. T. Næs O. Tomic and H. Martens, 'Path modelling by sequential pls regression', *Journal of Chemometrics*, vol. 25(1), pp. 28–40, 2011. DOI: https://doi.org/10.1002/cem.1357.

[21] E. Menichelli, T. Almøy, O. Tomic, N. V. Olsen and T. Næs, 'So-pls as an exploratory tool for path modelling', *Food Quality and Preference*, vol. 36, pp. 122–134, 2014, ISSN: 0950-3293. DOI: https://doi.org/10.1016/j.foodqual.2014.03.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095032931400055X.

[22] Q. C. Nguyen, K. H. Liland, O. Tomic, A. Tarrega, P. Varela and T. Næs, 'So-pls as an alternative approach for handling multi-dimensionality in modelling different aspects of consumer expectations', *Food Research International*, vol. 133, p. 109 189, 2020, ISSN: 0963-9969. DOI: https://doi.org/10.1016/j.foodres.2020.109189. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0963996920302143.

[23] *Python Machine Learning (Second Edition)*, eng. Packt Publishing, 2017.

[24] R. Fisher, *Iris data set*, data retrieved from UCI Machine Learning repository, https://archive.ics.uci.edu/ml/datasets/iris, 1936.

[25] V. Alto. (). Understanding the ols method for simple linear regression, [Online]. Available: https://towardsdatascience.com/understanding-the-ols-method-for-simple-linear-regression-e0a4e8f692cc. (accessed: 04.07.2020).

[26] S. Wold, M. Sjöström and L. Eriksson, 'Pls-regression: A basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001, PLS Methods, ISSN: 0169-7439. DOI: https://doi.org/10.1016/S0169-7439(01)00155-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169743901001551.

[27] K. Jöreskog and H. Wold, *Systems Under Indirect Observation: Causality, Structure, Prediction*, ser. Contributions to Economic Analysis nr. 139,poeng 2. North-Holland, 1982, ISBN: 9780444863010. [Online]. Available: https://books.google.no/books?id=Suq4AAAAIAAJ.

[28] F. Lindgren, 'The kernel algorithm for pls', *Journal of chemometrics*, vol. 7, pp. 45–59, 1993.

[29] P. H. Garthwaite, 'An interpretation of partial least squares', *Journal of the American Statistical Association*, vol. 89:425, pp. 122–127, 1994, ISSN: 0308-8146. DOI: 10.1080/01621459.1994.10476452.

[30] H. Wold, 'Partial least squares', in *Encyclopedia of Statistical Sciences*. American Cancer Society, 2006, ISBN: 9780471667193. DOI: 10.1002/0471667196.ess1914.pub2. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471667196.ess1914.pub2. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess1914.pub2.

[31] J. Wegelin, 'A survey of partial least squares (pls) methods, with emphasis on the two-block case', *Technical report*, Apr. 2000.

[32] J. A. Westerhuis, T. Kourti and J. F. MacGregor, 'Analysis of multiblock and hierarchical pca and pls models', *Journal of Chemometrics*, vol. 12, no. 5, pp. 301–321, 1998. DOI: 10.1002/(SICI)1099-128X(199809/10)12:5⟨301::AID-CEM515⟩3.0.CO;2-S. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-128X%28199809/10%2912%3A5%3C301%3A%3AAID-CEM515%3E3.0.CO%3B2-S.

[33] H. Lohninger, *Fundamentals of Statistics*. Epina eBook-Team, 2012. [Online]. Available: http://www.statistics4u.com/fundstat_eng/dd_pls.html.

[34] M. T. Alseth, 'Comparison of separate and joint modeling of bivariate response with emphasis on pls', Master's thesis, Norwegian University of Life Sciences, Aas, Norway, 2015.

[35] C. Cortes and V. Vapnik, 'Support-vector networks', *Machine Learning*, vol. 20, no. 273–297, 1995. DOI: 10.1007/BF00994018. [Online]. Available: https://doi.org/10.1007/BF00994018.

[36] W. Commons. (2018). Support vector machine. File: $\texttt{SVM}_{m}argin.png$, [Online]. Available: https://commons.wikimedia.org/wiki/File:SVM_margin.png.

[37] M. S. Beeri, J. Schmeidler, M. Sano, J. Wang, R. Lally, H. Grossman and J. M. Silverman, 'Age, gender, and education norms on the cerad neuropsychological battery in the oldest old', vol. 67, no. 6, pp. 1006–1010, 2006, ISSN: 0028-3878. DOI: 10.1212/01.wnl.0000237548.15734.cd. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163090/.

[38] O. Tomic, T. Graff, K. H. Liland and T. Næs, 'Hoggorm: A python library for explorative multivariate statistics', *The Journal of Open Source Software*, vol. 4, no. 39, 2019. DOI: 10.21105/joss.00980. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00980.

[39] T. E. Oliphant, *A guide to NumPy*. Trelgol Publishing USA, 2006, vol. 1.

[40] 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*, vol. 12, pp. 2825–2830, 2011.

[41] W. McKinney, 'Data structures for statistical computing in python', in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 51–56.

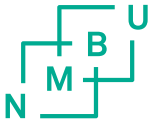
[42] S. Raschka, 'Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack', *The Journal of Open Source Software*, vol. 3, no. 24, Apr. 2018. DOI: 10.21105/joss.00638. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00638.

[43] H. I. Jacobs, M. P. Van Boxtel, J. Jolles, F. R. Verhey and H. B. Uylings, 'Parietal cortex matters in alzheimer's disease: An overview of structural, functional and metabolic findings', *Neuroscience and biobehavioral reviews*, vol. 297-309, pp. 134–147, 2012. DOI: https://doi:10.1016/j.neubiorev.2011.06.009. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21741401/.

[44] A. M. Brickman, J. Muraskin and M. E. Zimmerman, 'Structural neuroimaging in altheimer's disease: Do white matter hyperintensities matter?', *Dialogues in clinical neuroscience*, vol. 11(2), pp. 181–90, 2009. DOI: https://doi.org/10.31887/DCNS.2009.11.2/ambrickman. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864151/.

[45] F.-E. de Leeuw, F. Barkhof and P. Scheltens, 'Progression of cerebral white matter lesions in alzheimer's disease: A new window for therapy?', *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. 9, pp. 1286–1288, 2005, ISSN: 0022-3050. DOI: 10.1136/jnnp.2004.053686. eprint: https://jnnp.bmj.com/content/76/9/1286.full.pdf. [Online]. Available: https://jnnp.bmj.com/content/76/9/1286.

Thank you.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway