PROJECT 1: USING SUPERVISED LEARNING FOR CLASSIFICATION AND REGRESSION OF MRI DATA TO ESTIMATE CLINICAL DEMENTIA RATING

Jacob E. F. Overgaard (s144029) & Younes Subhi (s153611)

02450 Introduction to Machine Learning and Data Mining

1. INTRODUCTION (BOTH)

1.1. Dementia

Dementia is a general term for a broad range of neurodegenerative diseases that cause a gradual and lasting decrease in brain function, specifically in the cognitive functions of reasoning and memory. The decline in mental capabilities develop to a level that interferes with daily tasks and inhibit the plausibility of living independently. Alzheimer's is the most common type of dementia.

1.2. The problem of interest

Being able to foresee or predict the degree of dementia early, can help doctors improve treatment and delay the negative impact of dementia. This will not only help the diseased, but it will also help the health care system in improving quality of medication, and hopefully reduce the time to treatment. This report works with a data set governing features extracted from magnetic resonance images (MRI) of human brains. The goal is to classify the degree of dementia.

The data set, OASIS1 is obtained from "Open Access Series of Imaging Studies" (OASIS)[1] and is the first of three available brain imaging result data sets. OASIS works on making brain imaging results freely available to the public and present the findings in Journal of Cognitive Neuroscience [1].

Presented by Marcus et al. [1] the subjects' MRI images are used to estimate intracranial and brain volume. This is achieved by carrying out image analysis on MRI images, yielding physical attributes: estimated Total Intracranian Volume (eTIV), normalized Whole Brain Volume (nWBV) and an Atlas Scaling Factor (ASF). Additionally, generic assessments of health in the form of a 'Mini-Mental State Evaluation' (MMSE), Clinical Dementia Rating (CDR), age, gender, educational level (Educ) and socioeconomic status (SES) is registered for all test subjects. All MRI images are quality graded on a grade from 1-3, where 1 is best and 3 is worst. All MRI images graded 3 were originally removed from the results to maintain a high quality of test data. Bad quality MRIs are mostly caused by artifacts, electrical noise, poor head positioning or movement.

The findings presented by Marcus et al. [1] show the eTIV is good at automatically predicting the Total Intracranial Volume (TIV), opposed to earlier manual processes. Additionally the eTIV differ little with age and dementia rating. However, nWBV clearly shows a decline with aging accelerating at later life stages. In order to publicize the data, facial features of the scans have been removed, removing recognisable elements of the subjects. The authors claim that this do not effect the calculations of the anatomical measures of eTIV or nWBV.

If a classification of the MRI data is to be carried out, the interest is in determining whether a patient is likely to develop dementia within a certain time span. Thus CDR is the class label, and the primary attributes of interest are then Age, eTIV, nWBV, Educ, MMSE and SES. To solve the task, CDR is removed from the raw data and used as a class label of either 0, 0.5, 1 or 2. Additionally observations with missing or outlying data must be dealt with accordingly.

An alternative approach could utilize a multivariate regression for predicting nWBV based off of Age, Educ, MMSE and SES. This could be useful in approximating the nWBV without having to scan the brain and would be useful in the situation where a big number of subjects are to be investigated. For this operation the nWBV feature would need to be handled as the function $f(x_i)$ of the attributes x_i.

A straightforward alternative to the classification task could be clustering the data based on Age, eTIV, nWBV, Educ, MMSE and SES, similarly to the classification. In the clustering case it is of interest to see, if there may be new clusters of CDR, which could potentially do a more precise diagnosis of dementia.

The data set could additionally utilize the amount of discrete variables for association mining, by looking into association between attributes like Age, Educ, MMSE, SES and CDR. This could turn out helpful in early diagnosis of dementia and of specific social and educational classes of people.

The candidate is also a candidate for anomaly detection. An anomaly detection algorithm could be implemented, identifying anomalies in espicially eTIV or nWBV, based on the attributes Age, Educ, SES or MMSE. This could potentially

[1] URL: http://www.oasis-brains.org/. Accessed: 2019-05-02

be used in identifying faulty or erronous measurements and therefore be employed in ensuring a high quality MRI imaging results.

2. A DETAILED OVERVIEW OF THE DATA SET (BOTH)

The presented data set contains a total of 12 attributes measured on a total of 436 subjects. Three of the attributes are for this investigation removed. One is a test ID given to each MRI result. The second is a benchmarking test delay. The benchmarking test delay corresponds to the delay in days of an additional MRI scan, for a total of 20 test subjects. This delay is used for determining the reliability of the analytical approach. The delay attribute corresponds to 20 observations out of a total of 436 observations. The last disregarded attribute is the preffered hand of the test subject. All the observations are right-handed and this attribute can thus be disregarded.

The remaining 9 attributes and their corresponding types are given in Table 1.

Table 1: Data set attributes and their corresponding type

No.	Attribute	Description	Type
x_1	M/F	Gender (male/female)	Binary, Nominal
x_2	Age	Age in years	Discrete, Ratio
x_3	Educ	Education level (1 is lowest, 5 is highest)	Discrete, Ordinal
x_4	SES	Socioeconomic status (1 is highest status, 5 is lowest status)	Discrete, Ordinal
x_5	MMSE	Mini-mental state examination (30 is highest, 0 is lowest)	Discrete, Ordinal
x_6	CDR	Clinical dementia rating. 0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD	Discrete, Ordinal
x_7	eTIV	Estimated total intracranial volume cm^3	Continuous, Ratio
x_8	nWBV	Normalized whole brain volume	Continuous, Ratio
x_9	ASF	Atlas scaling factor (unitless)	Continuous, Ratio

2.1. Data Issues

Out of the 436 subjects, 216 subjects have data on all attributes. 201 subjects are missing data points on 'Educ' , 'MMSE' and 'CDR', while 220 subjects are missing data on 'SES'. No corrupted or extreme outlying data points is found in the dataset.

2.2. Statistical Summary

The table 2 gives a a description of the attributes, their value range and their mean value in the data set.

Table 2: Statistical Summary of the Attributes

Attribute	Range	Mean	σ
M/F	Male / Female	NaN	NaN
Age	18 – 96	51.36	12.3
Educ	1 – 5	3.18	3.2
SES	1 – 5	2.5	1.1
MMSE	14 – 30	27.06	3.4
CDR	0 – 2	0.02	0.38
eTIV	1123 – 1992	1481.92	160.5
nWBV	0.64 – 0.84	0.75	0.05
ASF	0.88 – 1.56	1.2	0.13

From table 2 it is clear, that most of the MRI subjects are reasonably old, in fact the average age is 51 years. This comes as no surprise, since dementia becomes more present with age. The standard deviation of the age is 12.3 years. The subjects come from all educational classes (Educ) and social classes (SES) with an average of 3.18 and 2.5 respectively. The mental health state as expressed by MMSE is fairly high with an average of 27.06 and a standard deviation of 3.4. The CDR, the attribute which we wish to classify, has a mean of 0.02. This specific mean does not entirely make sense, since it's a discrete and ordinal attribute. The mean of eTIV is 1481 with a standard deviation of 160. By a quick inspection of table 2, eTIV could appear to be normally distributed given its range of values and the corresponding mean. This is naturally only a quick estimate and histograms and boxplots are needed for a final conclusion. The normalized Whole Brain Volume (nWBV) has a mean of 0.75 and a standard deviation of 0.05. Like eTIV nWBV could appear to be normally distributed as well. ASF has a mean of 1.2 and a standard deviation of 0.13.

3. DATA VISUALIZATION (BOTH)

3.1. Data Investigation

An initial preprocessing step with larger data sizes is to investigate for any significant outliers in the data, which may effect

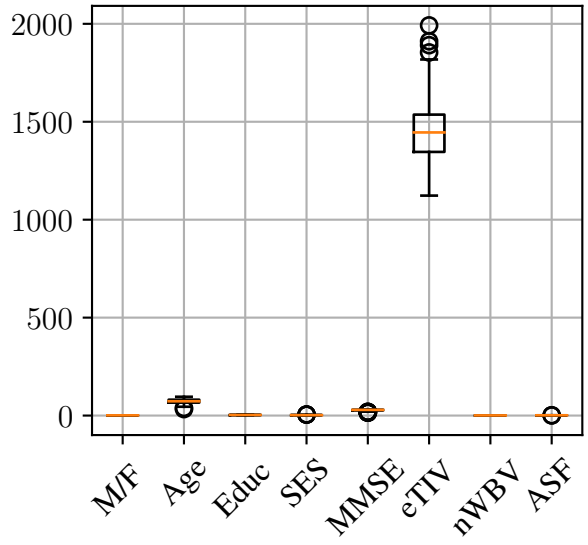

Fig. 1: Box plot visualization of each attribute included in the analysis of the raw OASIS data set.

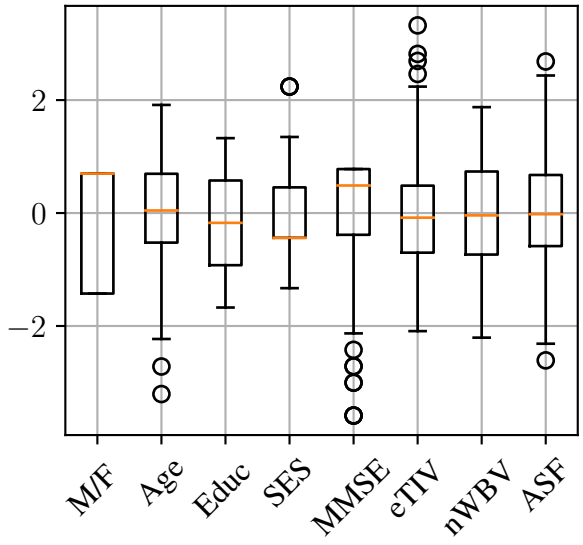

Fig. 2: Box plot visualization of each attribute included in the analysis of the standardized OASIS data set.

the results of the PCA analysis. One method of investigating outliers is to visualise the data using box plots.

While box plots can be a great tool of visualizing the basic statistics of a data set, plotting all the same attributes on the same scale, may give rise to issues due to different attributes potentially working on different scales. Because of this a standardisation of the data may be necessary in order to make the plots usable.

Standardizing the data transforms the data to be centered around zero-point. This makes it easier to spot outliers in the visualisation. In the visualisations CDR is removed, since it is the class label that the analysis tries to classify. In Figure 1 the raw un-standazised data is shown in a box plot. Due to the scale difference of e.g. eTIV, the box plot do not convey a substantial amount of information.

Figure 2 on the other hand is standardized and information on each attribute is easily accessed. Figure 2 shows that age, SES, MMSE, eTIV and ASF have outliers. Although due to either lack of substance or common pattern among the outliers, it would be wrong to exclude them.

A histogram can be used to investigate the distributions of data in the attributes. In Figure 3 the data distribution of the 8 attributes included in the analysis is visualised with histograms. While all feature distributions seem to be as expected for the data set, five of the eight features seem to be close to a normal distribution.

In order to investigate whether the attributes are independent or correlated, a correlation analysis is carried out, using a pair plot, which as the name suggests pairs all the attributes in all possible combinations. In Figure 4 a pair plot is shown for the attributes of the analysis. A correlation appears to be visible between eTIV and ASF of:

$$\mathrm{cor}\,(\mathrm{eTIV}, \mathrm{ASF}) = -0.99$$

This comes as no surprise giving the origin of ASF, that stems directly from the derivation of eTIV. However, between age and nWBV there is a correlation of:

$$\mathrm{cor}\,(\mathrm{age}, \mathrm{nWBV}) = -0.74$$

Which was also concluded by Marcus et al. [1]. Overall based on the initial data visualisations, the data set seems to be feasible for a machine learning based modeling.

3.2. Dimensionality Reduction - Principal Component Analysis

For doing a dimensionality reduction principal component analysis (PCA) is applied. Often data sets require several dimensions to portray the data. Reducing the amounts of dimensions required for handling the data often eases the entire process, both computer performance wise and data science wise. Additionally a PCA can show new trends or ways to classify or handle a given task.

However, when doing a PCA some aspects of the process must be kept in mind. Firstly reducing the dimensions with a PCA is not a lossless reduction. This in turn means, that if insufficient dimensions of the PCA is employed, the explained variance of the new subspace will distort the data. A PCA analysis creates a new subspace in which the original data can be projected based on the variation of the origin data.

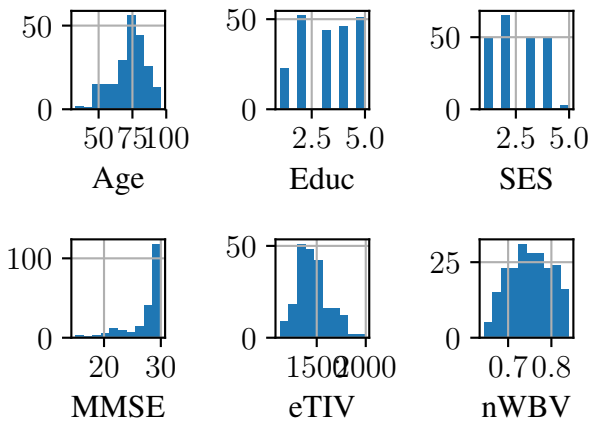

Fig. 3: Histograms visualizing data distribution of each attribute.

This means, if some attributes are vastly different in scaling from the others, the attributes must be normalized.

All observations with missing data points in at least one attribute, are removed from the sample for analysis (220 observations total) leaving 216 observations for the PCA. Based on the data visualisation, no further data manipulation is performed.

Let $\mathbf{X}$ be an $N \times M$ matrix of the OASIS1 data set, where N corresponds to the amount of observations and M correspond to the attributes. For now N and M are given by:

$$N = 216$$
$$M = 9$$

For the OASIS1 data set the different attributes are listed in Table 1. There are several different attribute types and their scaling is varying. The attributes Educ and SES are ordinal attributes ranging from 1 to 5, while eTIV is a continuous ratio attribute measuring intracranial volume in cm^3. Assuming the intracranial volume is not in the 1–5 cm^3 range (which is also apparent in the dataset, where eTIV's mean is $\mu = 1458\,\mathrm{cm}^3$), eTIV will dominate a PCA where the data in $\mathbf{X}$ is not normalized.

Furthermore the M/F (gender) attribute is designated as M/F which has to be turned into a numerical binary (0/1) representation for a PCA. However, carrying out a PCA on a gender attribute denoted with either 0 or 1 for either gender will favor the gender which is designated as a 1. So therefore the gender must be one-out-of-K encoded (synonomous to one-hot encoded).

Doing a one-out-of-K encoding on the gender will add an additional attribute since gender is now set by two columns: $[M\ F]$. This way, either gender is represented with a 1 in the given gender column, and therefore only a single of the two gender values can be a 1 (hereof the one-hot encoding). Because the main task is classifying CDR, it is removed from the PCA analysis and used as a class label vector y. This makes the entirety of the data available remain in the size given by:

$$N = 216$$
$$M = 9$$

Next the data set is normalized by subtracting each column by the column's mean and dividing with the column's standard deviation. This yields a normalized data set designated as $\boldsymbol{X_n}$. Mathematically this is expressed as given in (1).

$$\boldsymbol{x_{ij}} = \frac{\left(\boldsymbol{x_i} - \frac{1}{N}\Sigma_{i=1}^{N}\boldsymbol{x_i}\right)}{\sqrt{\frac{1}{N-1}\Sigma_{i=1}^{N}\boldsymbol{x_{ik}^2}}} \tag{1}$$

Once the data set is normalized the singular value decomposition (SVD) given in (2)

$$\boldsymbol{X_n} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{2}$$

where $\boldsymbol{V}$ corresponds to a set of M orthonormal vectors, here known as the principal components, which are a solution to the eigenvalue problem from (3).

$$\boldsymbol{X_n}\boldsymbol{V_i} = \lambda_i \boldsymbol{V_i} \tag{3}$$

The resulting eigenvalue λ_i for each principal component $\boldsymbol{V_i}$, correspond to the variance explained σ_i^2 of the given principal component. The SVD from (2) arranges all M eigenvalues in the diagonal of $\boldsymbol{\Sigma}$ and thus becomes an $M \times M$ diagonal matrix. The diagonal elements of $\boldsymbol{\Sigma}$ are given in (4).

$$\boldsymbol{\Sigma}_{diag} = [26.42\ 22.76\ 17.28\ 13.4\ 12.1\ 7.41\ 6.8\ 1.47\ 0.] \tag{4}$$

Knowing the diagonal elements of $\boldsymbol{\Sigma}$ the variance explained by the principal component projection can be estimated using the Frobenius norm of a given reconstructed matrix $\boldsymbol{X'}$ from a given matrix $\boldsymbol{X}$ as given in (5) [2]

$$\text{Variance Explained} = \frac{||\boldsymbol{X'}||_F^2}{||\boldsymbol{X}||_F^2} = \frac{\Sigma_{i=1}^{n}\sigma_i^2}{\Sigma_{i=1}^{M}\sigma_i^2} \tag{5}$$

where M designates the M dimensions of the original data set $\boldsymbol{X}$, here $M = 9$ and n is the dimension of the projected subspace. Plotting (5) based on (4) gives the variance explained for the normalized OASIS1 data set $\boldsymbol{X_n}$ in Figure 5 with a 0.9 threshold.

As is apparent from Figure 5 it takes at least five principal components to reach the threshold of 0.9. The threshold in this explained variance is somewhat arbitrary and highly application specific. The first principal component accounts for 0.36 of the total variance explained. This indicates that the

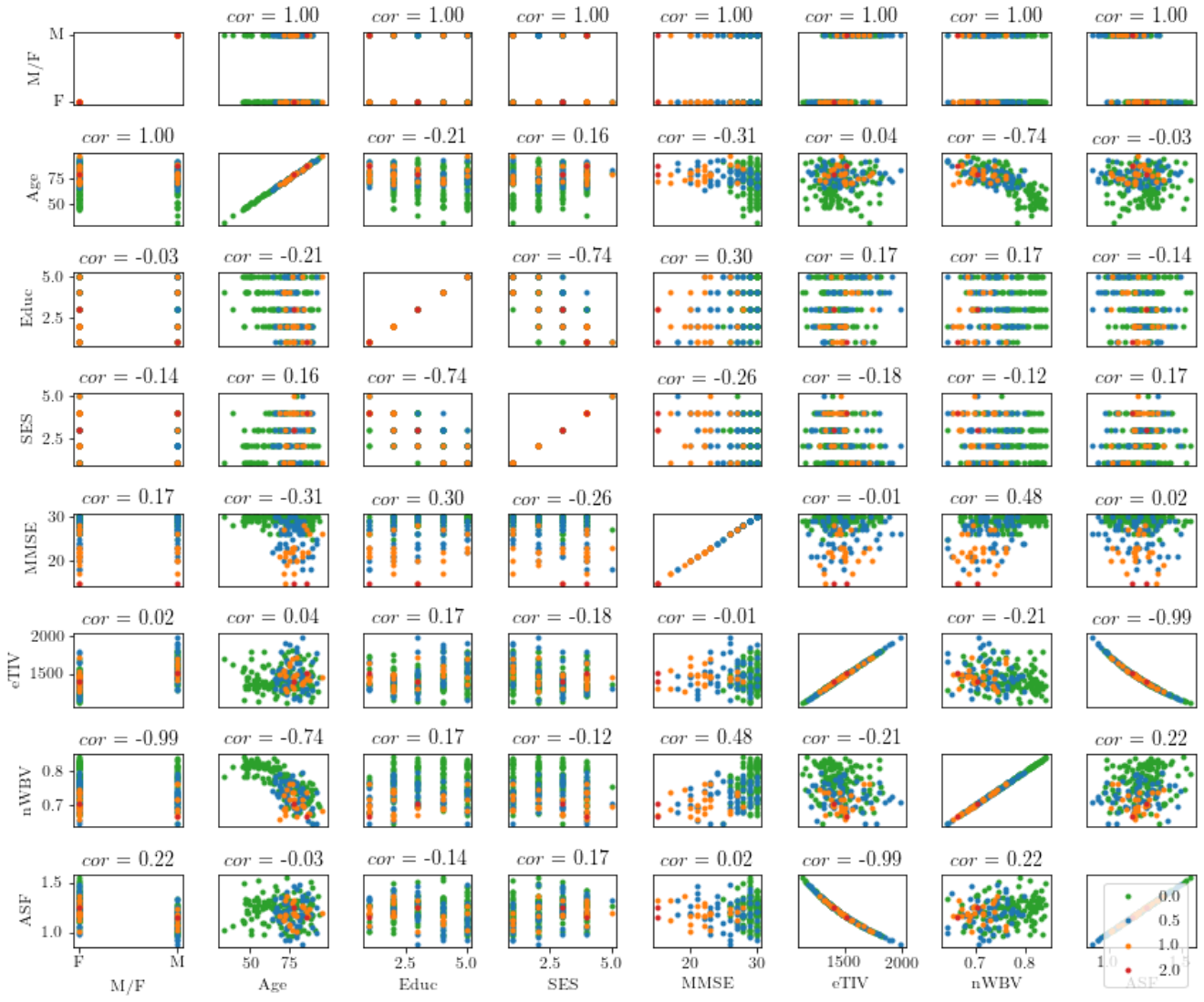

Fig. 4: Pair plot visualizing the correlation of all combinations of the 8 features used in the analysis.

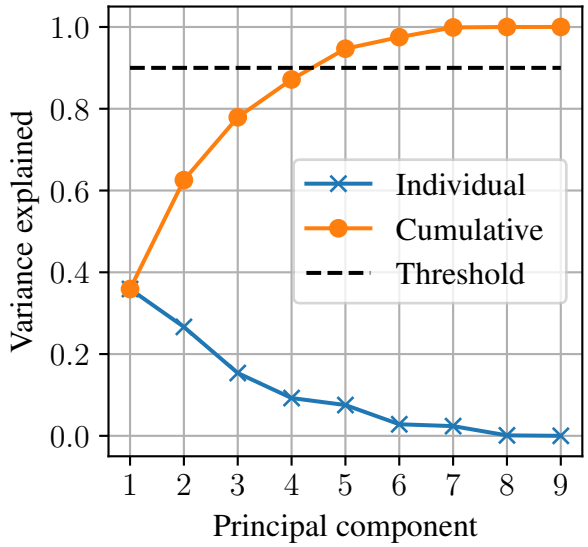

Fig. 5: Cumulative and individual variance explained by principal components in $\boldsymbol{V}$ with a threshold of 0.9. A minimum of five principal components are required for explaining more than 90% of the variance.

original normalized data $\boldsymbol{X_n}$ is not able to be projected onto a 1-dimensional subspace without losing 67% of the original variance. Similarly the first two components correspond to 0.63 of the variance explained. To increase the variance explained for the first two principal components. The variance of the first two principal components ideally explain how much, the original data fits onto a two-dimensional plane.

When it comes to the principal components of $\boldsymbol{V}$, only the first three principal components are included due to space restriction. For a more detailed view of the principal components, we refer to the PCA.py appendix. The three first principal components from $\boldsymbol{V}$ are given in (6).

$$\boldsymbol{V_{123}} = [\boldsymbol{V_1}\ \boldsymbol{V_2}\ \boldsymbol{V_3}] = \begin{bmatrix} -0.05 & -0.12 & 0.12 \\ 0.46 & -0.44 & 0.41 \\ -0.43 & -0.47 & 0.52 \\ 0.04 & 0.23 & -0.19 \\ -0.47 & 0.09 & -0.12 \\ 0.01 & 0.71 & 0.7 \\ -0.62 & -0.04 & 0.02 \\ 0.01 & 0.02 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{6}$$

To analyze each individual attribute's effect onto the subspace spanned by $\boldsymbol{V_1}$ (PC1), $\boldsymbol{V_2}$ (PC2) and $\boldsymbol{V_3}$ (PC3), the principal component coefficients based on attributes are shown in Figure 6.

The first thing to note from Figure 6 and (6) is, that neither of the three first principal components account for the female gender. In fact only PC8 and PC9 project the female gender onto the subspace. Additionally, the first two principal components project very little of the male gender, and the third principal component none of the male gender.

Looking into the first principal component, Age, SES, eTIV and ASF all have a negative influence on the principal component. This means if either of those attributes are below the corresponding mean, they will have a net positive influence. Educ, MMSE and male gender all have a positive projection onto the first principal component. The primary attributes projected by PC1 are ASF and eTIV, followed by Educ and SES.

The second principal component primarily projects nWBV, followed by SES and Educ. The attributes nWBV and MMSE both have a positive projection onto the second principal component. This means for nWBV and MMSE above their corresponding mean, they will have a net positive projection. Age, Educ, SES and ASF all have negative projections onto PC2.

The third principal component positively project nWBV, SES, EDuc Age and ASF in descending order. MMSE and eTIV are both projected negatively onto PC3.

To successfully use the possibilities of dimensionality reduction, the data set $\boldsymbol{X_n}$ must be projected utilizing the new orthonormal basis given by $\boldsymbol{V}$. The projection $\boldsymbol{B}$ of $\boldsymbol{X_n}$ onto a subspace $\text{span}\,(\boldsymbol{V_1}, \boldsymbol{V_2})$ is given by (7).

$$\boldsymbol{B} = [\boldsymbol{b_1}\ \boldsymbol{b_2}] = \boldsymbol{X_n}\,[\boldsymbol{V_1}\ \boldsymbol{V_2}] \tag{7}$$

The resulting projection is shown in Figure 7.

From Figure 7 it is clear that there is correlation between $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ of the projected data. In fact:

$$\text{cor}\,(\boldsymbol{b_1}, \boldsymbol{b_2}) = -0.99$$

where the negative sign is expected since for increasing PC1, PC2 is decreasing. From the 2-dimensional results there is no apparent way of classifying either rating of CDR. Although, it seems for increasing PC1 the likelihood of worse degrees of dementia becomes more plausible. It must be taken into consideration, that for the same areas of worse degrees of CDR, healthy subjects (green dots) are observed.

It comes as no surprise, that a 2-dimensional projection lacks a lot of information due to the variance explained of 0.63 for the first two principal components. An alternative approach is to increase the dimension of the subspace from $\mathbb{R}^2$ to $\mathbb{R}^3$ and see, if enough variance is then expressed.

A 3-dimensional projection as given in (8) is shown in Figure 8.

$$\boldsymbol{B} = [\boldsymbol{b_1}\ \boldsymbol{b_2}\ \boldsymbol{b_3}] = \boldsymbol{X_n}\,[\boldsymbol{V_1}\ \boldsymbol{V_2}\ \boldsymbol{V_3}] \tag{8}$$

From Figure 8 the same tendency with correlation is observed. The correlations between the first three dimensions of

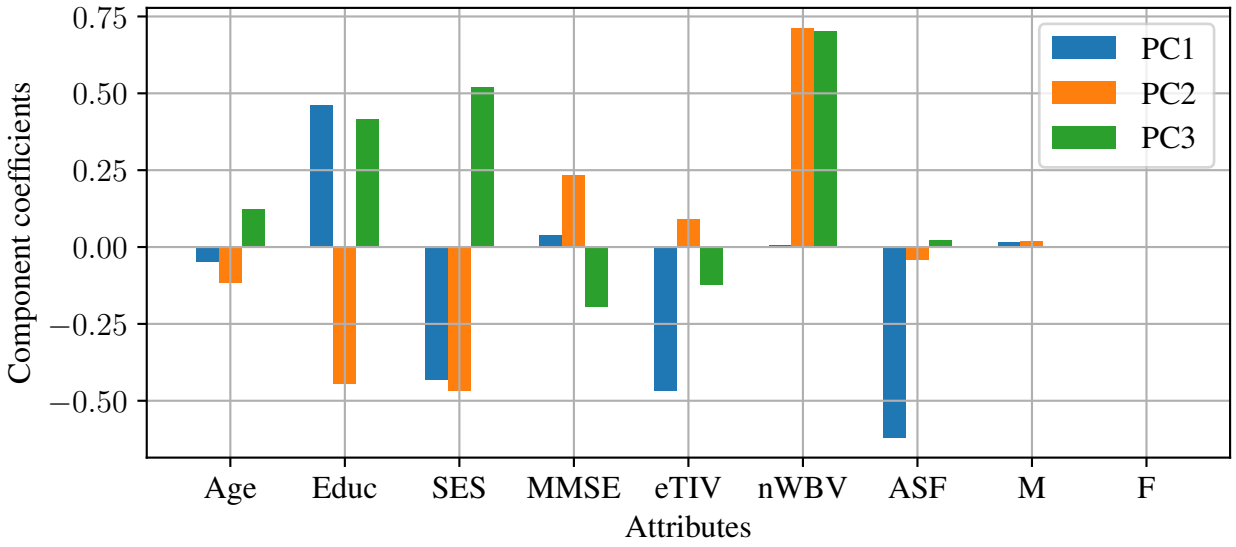

Fig. 6: Principal component coefficients based on $\boldsymbol{X}$'s attributes for $PC1$, $PC2$ and $PC3$.

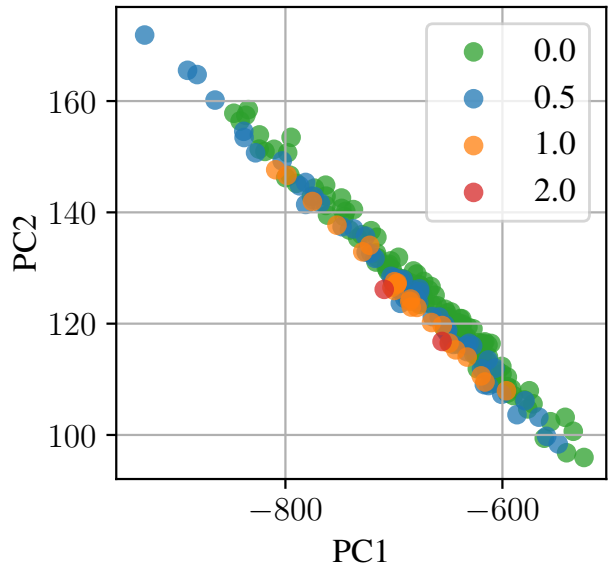

Fig. 7: A projection of $\boldsymbol{X_n}$ onto $\mathbb{R}^2 = \boldsymbol{X_n}$ onto span $(\boldsymbol{V_1}, \boldsymbol{V_2})$ for different classes of CDR.

$\boldsymbol{B}$ can be given as:

$$\begin{aligned} \text{cor}\,(\boldsymbol{b_1}, \boldsymbol{b_2}) &= -0.99 \\ \text{cor}\,(\boldsymbol{b_1}, \boldsymbol{b_3}) &= 0.99 \\ \text{cor}\,(\boldsymbol{b_2}, \boldsymbol{b_3}) &= -0.99 \end{aligned}$$

While using a 3-dimensional projection of the data yields a much better result than Figure 7, there is still no entirely clear boundary between each of the four CDR classes. Increasing the subspace span by $\boldsymbol{V}$ naturally increases the likelihood of seeing any patterns, but portraying more than three dimensions is not feasible.

So to clearly classify between the different classes of CDR, further investigations must be done in higher dimensions and further in the work.

4. CONCLUSION (BOTH)

The data from OASIS data set consists of 12 attributes, of which 3 have been discarded in this analysis. Of the remaining 9 attributes, a single is binary and the rest numerical. 220 observations had missing attributes and have been removed, leaving 216 for the analysis. The binary attribute (gender) is one-out-of-K encoded to ensure no favouring, and the rest of the attributes are normalised for carrying out a principal component analysis. If the data is not normalized and binary attributes like gender is not one-ouf-of-K encoded, the principal component analysis changes drastically. Additionally handling a classification class, this one must also be left out, in order for it not to distort the data.

From the data visualization it is learned that most of the attributes do not seem directly correlated with one another. Furthermore it is concluded that ASF, Age, nWBV and eTIV can be regarded as normally distributed.

While the principal component analysis went through the general procedure of dimensionality reduction, clearified the loss of variance and inspected the new subspace's orthornormal basis, a PCA analysis could not directly visualize a clear solution to the classification task at hand. Therefore further investigation must be performed.

5. REFERENCES

[1] Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner, "Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, September 2007.

[2] Tue Herlau, Mikkel N. Schmidt, and Morten Moerup, *Introduction to Machine Learning and Data Mining*, Number 1. January 2019.

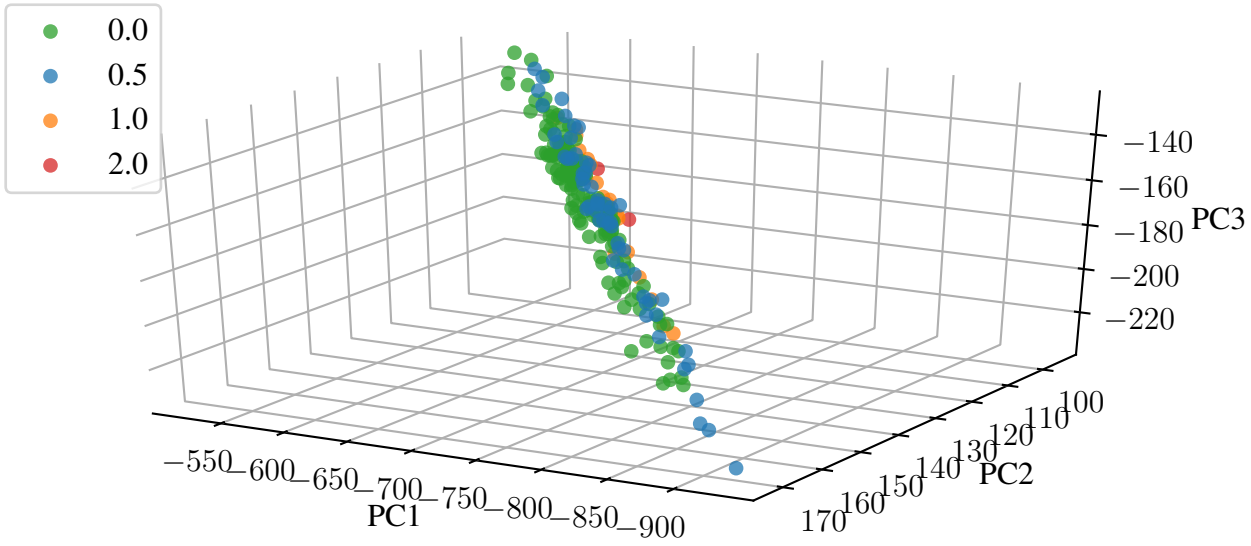

Fig. 8: A projection of $\boldsymbol{X_n}$ onto $\mathbb{R}^3 = \text{span}\,(\boldsymbol{V_1}, \boldsymbol{V_2}, \boldsymbol{V_3})$ for different classes of CDR.