Predicting dementia status

*How accurate can a machine learning model be, that predicts if a subject has dementia using different clinical parameters?*



Ewoud Ooms

387577

BFV3

5 oktober 2023

## Abbreviations/Codebook

**Afbeelding met tekst, schermopname, Lettertype, menu

Automatisch gegenereerde beschrijving**

Introduction

Dementia is a pressing global health concern, with a significant impact on individuals, families, and healthcare systems. Timely diagnosis and intervention are crucial for improving the quality of life for those affected by dementia. Advances in machine learning and healthcare technology offer promising opportunities to enhance the accuracy and efficiency of dementia diagnosis.

The question this research is aiming to give an answer to is:

How accurate can a machine learning model be, that predicts if a subject has dementia using different clinical parameters?

Our approach combines machine learning and dementia research to uncover hidden patterns in clinical data. We will conduct an Exploratory Data Analysis (EDA) to identify correlations with the dementia group, assisting in feature selection and model development.

Dataset: https://www.kaggle.com/datasets/shashwatwork/dementia-prediction-dataset

## **Results**

**Finding variation and looking for coherence**

The things to look for in a boxplot is to see if the values of the dementia groups are far apart of each other with the parameters, because this means that the influence of the parameters effects the groups different and is therefore maybe a good parameters for correlation and the machine learning model

Afbeelding met tekst, diagram, schermopname, Plan

Automatisch gegenereerde beschrijvingAfbeelding met tekst, diagram, schermopname, Plan

Automatisch gegenereerde beschrijvingAfbeelding met diagram, tekst, Plan, lijn

Automatisch gegenereerde beschrijving

By eyeballing the following figures, we can conclude the following

The parameters with the most difference are : EDUC, SES, MMSE and nWBV

These parameters are mostly the ones that are going the be used to make a model.

**Finding correlation**

A way to find correlation within the data set is to use an anova test. We will use an anova test to see which parameter is correlated the most with the dementia group. The anova test uses the eta-squared to show the correlation.

Eta-squared ranges from 0 to 1 and is interpreted as follows:

η² = 0: There is no effect of the independent variable on the dependent variable.

η² ≈ 0.01: A small effect.

η² ≈ 0.06: A medium effect.

η² ≈ 0.14: A large effect.

The data is normally distributed except for MMSE so this test is just a ranking.

Afbeelding met tekst, schermopname, diagram, Perceel

Automatisch gegenereerde beschrijving

When looking at the outcome we see that MMSE, nWBV, EDUC and Age have an effect on the dementia status.

We will be taking these parameters for further analysis.

Another way to check for correlation is with a correlation matrix that is shown in a heat map.

Using a heat map is a great way to visually represent the correlations between different parameters with colors.

Afbeelding met tekst, schermopname, nummer, Kleurrijkheid

Automatisch gegenereerde beschrijving

A positive correlation means that both variables increase or decrease together. A negative correlation means that one variable increase while the other variable decreases.

The things we are looking for are surprisingly high correlation numbers and the correlation numbers between the parameters and the dementia group.

We can see that the correlation number between ASF and eTIV is almost -1. This means that the eTIV value and the ASF value are almost entirely coherent to each other and that we can delete one of the other because they say the same thing.

The groups with the highest correlation number with the dementia group are the MMSE, EDUC, SES, and nWBV parameters.

These parameters were also very high in the anova test, except for the SES parameter.

Using the Pearson's Chi-squared test we tested that the p value is below 0.05 and is therefore correlated with the dementia status.

**Finding clusters**

So, we now have a clear view of the parameters that are highest correlated with the dementia group.

Before we are going to the machine learning process, we are going to check how good the model is probably going to be.

The use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) to observe trends, jumps and clusters. This overview may uncover how good of a model we can make.

Looking for clusters with a principal component analysis plot using the 4 most correlated parameters with dementia.

MMSE, EDUC, nWBV and SES

Afbeelding met tekst, diagram, schermopname, Plan

Automatisch gegenereerde beschrijving

Clearly 3 cluster groups can be seen, demented and non demented lie nicely apart with little overlap and the converted group sits as a clear middle group in between with more similarity to non demented as the article had already indicated. What can be gleaned from this is that 3 groups are with clearly different values so making a machine learning model to predict the 3 groups is probably quite possible.

With the PC1 and PC2 component there is 72 % variation which means is that with 2 parameters 72 % of the data can be correctly identified.

This means that the accuracy of our model will probably be around that number.

If we have a model that has 99 % accuracy, we are probably doing something wrong because the PCA plot is showing us that that won't be possible.

Discussion and conclusion

Discussion

My General view I have of the quality of the data set I used in this research is that the data set was very good and very clean and almost had no missing data. The dataset is a little small but big enough to have proper research.

The possibilities were great with this dataset, I had to change some parameters from character to numeric but every plot or step that I wanted to take I could do.

A little problem with the dataset was that not every patient had the same number of visits and therefore not the same number of instances. I did not think it was a problem because the number of instances with dementia and without were almost the same so I don’t think it had an influence, but it would be nice to have the same amount of data for every patient.

Conclusion

After this EDA I think I have collected a clean version of the dataset and that I can go on to the next step and can start with the machine learning process.

There is a clear view of the parameters with the high correlation to the dementia status and I have high hopes that I can make a good model.

## References

Code book:

Ruslankl. (2018). Dementia Prediction w/ Tree-based Models. *Kaggle*. <https://www.kaggle.com/code/ruslankl/dementia-prediction-w-tree-based-models>

Data:

Battineni, G., Chintalapudi, N., & Amenta, F. (2019). Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked*, *16*, 100200. <https://doi.org/10.1016/j.imu.2019.100200>