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

12 November 2023

**Abbreviations**

The codebook [1] with all the attributes from the dataset and there abbreviation, for extra explanation see Thema09DementiaPrediction.docx [3]

**Afbeelding met tekst, schermopname, Lettertype, menu

Automatisch gegenereerde beschrijving**

**Table of Contents**

**1. Introduction ................................................. [3]**

**2. Materials and Methods ........................................ [4]**

**3. Results ...................................................... [6]**

**4. Discussion ................................................... [12]**

**5. Conclusion ................................................... [13]**

**6. References ................................................... [14]**

**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>

**Materials and methods**

This section will go over the materials and methods used to set up the tool are discussed.

**Materials**

To filter the data and determine which attributes are most correlated with the dementia status, an Exploratory Data Analysis (EDA) was conducted. For this purpose, RStudio, utilizing the R programming language, was employed [5].

The packages downloaded and utilized for this project are documented in the file Thema09DementiaPrediction.docx [3].

To construct the model and ascertain which algorithm will get the best accuracy, data mining software Weka was employed [6].

**Methods**

1. **Data processing:**

Prior to analysis, several preprocessing steps were undertaken, and later subsets were made for model making.

- **Attribute deletion:**

At the start of the project 5 attributes were deleted from the dataset. The MRI I, delay and visit because they have no influence on the dementia status. CDR because it is basically a parameter telling if a patient has dementia or not. Handedness because every patient is right-handed, so it does not matter.

- **Conversion of Converted Group:**

The initial dataset included a "Converted" group, representing patients initially classified as demented but later reclassified as non-demented. To streamline the analysis and align with the research focus, before the model making the "Converted" group was reassigned to the "non-demented" category. [2]

1. **Statistical Analysis:**

**- Statistical Tests:**

Statistical tests, such as the Anova test were conducted to find correlation within the data set using the Eta-squared score to rank the attributes.

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.

Also, a Pearson's Chi-squared test to check the p value of the SES score with the dementia status.

1. **Model Building:**

**- Attribute Selection:**

Weka attribute selection was conducted to identify attributes with high correlation to dementia status. The 'Select attributes' function in Weka was employed, involving two tests: one using 'cfsSubsetEval' with the 'Best first' method and another using 'WrapperSubsetEval' with the 'Best first' method and the J48 classifier. The selected attributes for the final model were determined based on the results of these tests. These were the MMSE, SES and nWBV attribute

**- Algorithms:**

- Various machine learning algorithms were employed to construct predictive models. These algorithms include:

|  |
| --- |
| 1. ZeroR (weka.classifiers.rules.ZeroR) |
| 2. OneR (weka.classifiers.rules.OneR) |
| 3. J48 (weka.classifiers.trees.J48) |
| 4. RandomForest (weka.classifiers.trees.RandomForest) |
| 5. IBk (weka.classifiers.lazy.IBk) |
| 6. CostSensitiveClassifier (weka.classifiers.meta.CostSensitiveClassifier) |
| 7. SMO (weka.classifiers.functions.SMO) |
| 8. SimpleLogistic (weka.classifiers.functions.SimpleLogistic) |
| 9. NaiveBayes (weka.classifiers.bayes.NaiveBayes) |
| 10. Bagging (weka.classifiers.meta.Bagging) |

**- Cross-Validation:**

- A 10 folds cross-validation was employed to assess the generalizability of the models.

**- Evaluation Metrics:**

- The performance of each model was evaluated using common metrics such as accuracy, precision, True positive/ False negatives, and the Area Under the ROC Curve.

## **Results**

**Finding variation and looking for coherence**

The initial key findings that emerged during the Exploratory Data Analysis (EDA) were observed while testing whether the attributes were associated with dementia status. As seen in the figure below, certain attributes immediately stand out.

The thing 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 parameter for correlation and the machine learning model

Afbeelding met tekst, schermopname, diagram, Lettertype

Automatisch gegenereerde beschrijving

Figure 1, this figure shows boxplots of the attributes with the dementia status groups to look for relation/correlation. In blue is the converted group, in red is the demented group and in green is the non-demented group.

Taking a good look at the figures above, we can spot some standouts. The ones that really stand out are the EDUC, SES, MMSE, and nWBV parameters. These probably will play a big part in the building of the model.

**Finding correlation**

Because of the boxplots we know that there probably is a significant relation between the attributes and the dementia status, for making a model we need to know which attributes have the biggest relation.

Another way to find relation within the data set is to use an Anova test. We will use an Anova test to see which parameter has a big relation with the dementia group. The anova test uses the eta-squared to show the correlation.

**Anova test of the attributes with the dementia status \*figure 2\***

Figure 2, an anova test to find see which parameter has the biggest relation to the dementia status.

Afbeelding met tekst, schermopname, diagram, Perceel

Automatisch gegenereerde beschrijving

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

We will be taking these parameters for further analysis.

**Heatmap with pair plot**

Afbeelding met tekst, schermopname, nummer, Kleurrijkheid

Automatisch gegenereerde beschrijvingNow, here's another way to check for correlations – Figure 3 shows the pairwise correlations between attributes that we observed as a correlation matrix by means of a heatmap

Figure 3, a heatmap to see how all the attributes relate to one another and to see correlation with the dementia status.

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.

Because the Anova test shows us that the relation between the SES parameters and the dementia status were low but the heatmap gives us a high number we use an extra test

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.

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

Afbeelding met tekst, diagram, schermopname, Plan

Automatisch gegenereerde beschrijvingMMSE, EDUC, nWBV and SES

Figure 4, A PCA plot to see the clustering of the 3 dementia status groups with the parameters that have the biggest relation.

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.

**Weka model**

Now that all results regarding associations and correlations have been obtained from the EDA, it's time to craft the machine learning model. All the steps made to get to these final findings can be found in the Thema09DementiaPrediction.docx [3]. This is the conclusionAfbeelding met tekst, schermopname, diagram, lijn

Automatisch gegenereerde beschrijving

Figure 5, a bar plot showing the accuracy of the different machine learning algorithms that have used to get the best model.

Looking at the accuracy of the classifiers it becomes clear that the difference in accuracy is not significantly enough between algorithms that the final model must be picked on something else. On closer inspection, the classifiers also don’t have a significant difference in False negatives. This is an important factor because it is no good if people with demented are wrongly classified.

The remaining factor is simplicity, the J48 model only uses the MMSE score to make its model. With only 1 attribute the model is simple to use and to understand. Let’s look at the ROC curve to make a final decision.

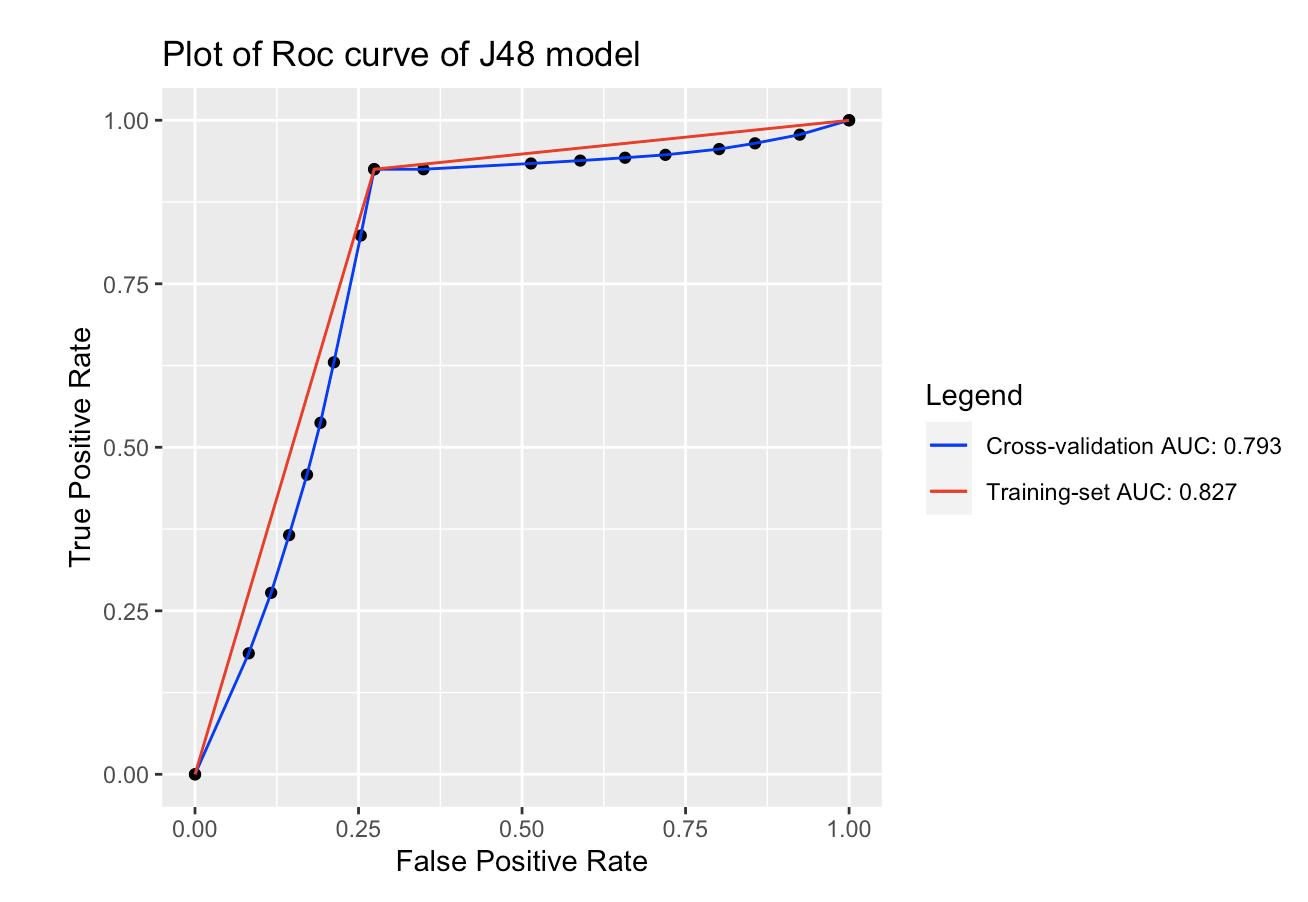


Figure 6, An ROC curve of the j48 model that will be used

The above ROC plot illustrates the performance of the J48 algorithm, evaluated through cross-validation and with the training set. A positive aspect is that the ROC scores of the two lines are closely aligned, indicating minimal overfitting. Although the score doesn't reach perfection (1), the J48 algorithm emerges as the highest-scoring algorithm for this dataset.

Discussion and conclusion

Discussion

My general view of the quality of the data set used in this research is that the data set was very good and very clean and almost had no missing data. One thing that stood out was the size difference between the Non-demented and Demented groups in the dataset. It probably did not play a big part but having a more balanced dataset would've been better.

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.

The final model only taking the MMSE score is a nice simply model but also brings its limits by only making two groups. There were a lot of attributes that were discarded at the beginning because they had little use, would like to have had some more attributes that maybe would have let to a bigger and better model.

**Conclusion and future work**

The aim of this study was to answer the question: "How accurate can a machine learning model be in predicting if a subject has dementia using different clinical parameters?".

I successfully built a model with a single attribute, achieving an accuracy of approximately 85%, with a specific focus on minimizing false negatives. Hence, I believe that the aim of this study is achieved but also think there is loads of room for improvement.

Maybe with some better attributes of patients and a bigger dataset there can be made a way more accurate model that can predict the status of dementia and could be implemented at nursing homes to maybe catch the dementia disease early on and start the medication for a slower decaying process of the mind and body.

## References

[1] Code book:

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

[2] 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>

[3] RStudio\_files:

<https://github.com/eaooms/Thema09-2023/tree/main/RStudio_Files>

[4] JavaWrapper\_Files:

<https://github.com/eaooms/Thema09-2023/tree/main/WekaAplicatie>

[5] Rstudio download:

<https://posit.co/download/rstudio-desktop/>

[6] Weka download:

<https://sourceforge.net/projects/weka/>