Logistic regression and decision trees for dementia prediction

1 Introduction & problem formulation

Dementia is a syndrome that leads to progressive deterioration in cognitive function beyond what is normally expected as a result of biological ageing. As there is no cure for dementia, preventative measures are extremely important. Identifying at-risk individuals is thus very important, and in this project I will investigate how machine learning can be used to predict susceptibility to dementia.

The data [2] used for this project is from a 2019 study [3] on using ML methods for dementia prediction, specifically support vector machines. In this project alternative classification methods will be studied. The data set consists of data from 373 separate imaging sessions administered on subjects aged 60 to 96. Each subject was administered a MRI scan on two or more visits, and the total number of participants is 150. Observe that a data point is a single MRI imaging of an individual, so we have 373 data points in total. As the subjects have been administered a MRI scan on two or more occasions, they appear in the dataset at least twice, but the imaging sessions are treated as individual data points.

The dataset, feature selection, ML methods and construction of the training, validation, and test sets are discussed in section 2. The results of our methods will be presented in section 3, and conclusions in section 4.

Summary of the problem

Datapoint: a single MRI imaging with background data of a 60-96 year old individual. 373 data points in total.
Label: is patient demented of non-demented at the time of the MRI imaging (label space $\mathcal{Y} = \{0, 1\}$).
Features: Age, gender, years of education, socioeconomic status (1-5, SES), clinical dementia rating (0-3, CDR), mini mental state examination (0-30, MMSE), estimated total intracranial volume (ETIV), normalized whole brain volume (NWBV) and Atlas Scaling Factor (ASF) (feature space $\mathcal{X} = \mathbb{R}^n$).

2 Methods

Data and preprocessing

The data is described above, but the definitions of some of the features might be unclear. The clinical dementia rating (CDR) is a 0–3 point numeric scale derived from clinician rating of cognition and daily function [4]. The largest CDR-value in this dataset is 2. The Mini Mental State Examination (MMSE) is a test designed as a screening test for the purpose of evaluating cognitive impairment in older adults [5]. Total intracranial volume is defined as the volume within the cranium [5], and the Atlas Scaling Factor (ASF) is a one-parameter scaling factor that allows for comparison of the estimated total intracranial volume (eTIV) based on differences in human anatomy [5]. The motivation behind using the classification of nondemented and demented patients as the label is described in the problem formulation: it is very important to identify at-risk individuals and also identifying causes of dementia.

As the feature values have very different ranges, the features will be normalized using the scikit-learn [7] StandardScaler -function, which standardizes features by removing the

the dataset mean of the feature and scaling to unit variance. Using StandardScaler to scale the data yielded better validation accuracies for both models than the MinMaxScaler, which was the motivation behind choosing this method of scaling the data before fitting.

Feature selection

Feature selection is done here from a very pragmatic standpoint, with the aim of finding out whether dementia can be predicted from an individuals medical history and background. With this in mind, we leave out the CDR (clinical dementia rating) from our model, as the CDR test is one of the main methods of diagnosing dementia, so it would partly defeat the purpose of predicting the label with a ML method. The rest of the features will be included in our model.

First method

The first ML method used is logistic regression. Logistic regression is a commonly used binary classification method, that uses logistic loss to measure the usefulness of a linear hypothesis $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. Thus the hypothesis space is the linear hypothesis space $\mathcal{H}^{(n)}$. The method tries to find the optimal parameter vector $\hat{\mathbf{w}}$ that minimizes the average logistic loss

$$\hat{L}(\mathbf{w}|D) = \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y^{(i)}h^{(w)}(\mathbf{x}^{(i)}))$$

for a labelled dataset $D = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{m}$ with m feature vectors $\mathbf{x}^{(i)}$ and labels $y^{(i)}$ [6, Chapter 3.6, Chapter 2.12]. Logistic loss is chosen as our loss function, as it allowed the use of the ready-made scikit-learn logistic regression classifier, and average zero-one loss will be used for computing the validation error in order to choose between our first and second method. I chose logistic regression as the first ML method, as the absolute value of the prediction $h(\mathbf{x})$ quantifies the confidence of the classification, which is very useful in this context, as in practice it is helpful to be able measure the risk of dementia of a patient instead of a binary classification.

Second method

The second ML method used is a decision tree with maximum depth $d = 2$. $d = 2$ was chosen based on examining model performance for $d = \{1, 2, \ldots, 14\}$. The model started to overfit very quickly for $d \geq 5$. I used scikit-learn's decision tree classifier with Gini impurity, which measures the quality of an individual decision region created by the model. When using a decision tree model for classification, some form of impurity is a common choice for our loss function [6, Chapter 3.10], and Gini impurity is the standard choice for the scikit-learn function. Again, average zero-one loss will be used for computing the validation error.

For both methods, in addition to the accuracy scores, recall and F1 scores will be used to measure performance, with the recall and F1 scores being calculated for each cross validation split and taking the averages. The F1 score is the harmonic mean of precision and recall, with precision defined as $\frac{true\ positive}{true\ positive+false\ positive}$ and recall as $\frac{true\ positive}{true\ positive+false\ negative}$. In this application, we specifically focus on recall rather than precision, as a false negative classification is potentially far more risky than a false positive.

Training and validation sets

In constructing our training and validation sets for fitting both models to the data, we will use k-fold cross validation. This is due to the fact that we have a relatively small dataset, so using a single split for the training and validation data might result in an unlucky split that isn't very representative of the data. The effect of such a split can be reduced by using k-fold cross validation [6, Chapter 6.2.2]. We will use 6-fold cross validation, which results in 6 different train-test splits, with each validation set being one sixth of the whole dataset for each fold. The results (discussed and presented in stage 3 submission) indicate that $k = 6$ seems to be a reasonable split that manages to average out the effects of unlucky splits. Before using 6-fold cross validation for our models, we take a test set of size 0.2 to use in our final comparison of the performance of our models.

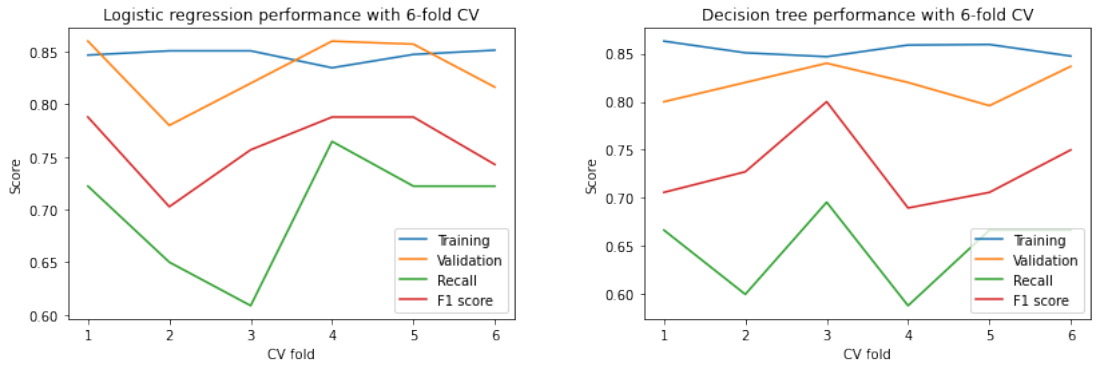
3 Results

The averages of performance measures over cross validation for our classification models are presented in table 1, and the performance measures for each cross validation fold in figure 1. Logistic regression slightly outperformed the decision tree model over the cross validation, especially regarding the recall score. Both models seemed to slightly overfit, but not dramatically. Based on this, the logistic regression model is the final chosen method, with 0.88 accuracy over the test set. After choosing the logistic regression as the final model, the performance of our decision tree model on the test set was also examined. Performance on the test set is presented in table 2 and figure 2.

Performance over cross validation:

Average:	Training accuracy	Validation accuracy	Recall	F1 score
Logistic regression	0.847	0.832	0.700	0.761
Decision tree	0.854	0.819	0.647	0.730

Table 1: Performance of models (averages over 6-fold validation)



(a) Logistic regression model over 6-fold CV

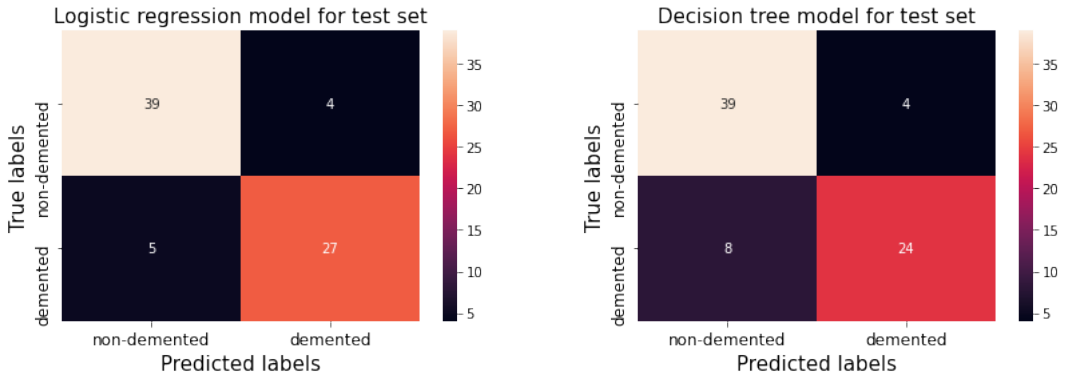
(b) Decision tree model over 6-fold CV

Figure 1: Model performance in cross validation

Performance over test set:

Test set	Accuracy	Recall	F1 score
Logistic regression	0.88	0.87	0.85
Decision tree	0.84	0.75	0.80

Table 2: Performance for test set



(a) Confusion matrix of LR model on test set (b) Confusion matrix of DT model on test set

Figure 2: Model performance on test set

4 Conclusion

Although the results of the logistic regression model are somewhat promising, there is clearly room for improvement, especially regarding the low recall scores. It would be interesting to examine the absolute values of the predictions of the falsely classified data points, to see if the model was "close" to classifying them correctly. Looking into what is the least number of easy to access features the model needs for adequate performance would be helpful for real-world applications, as the features derived from the MRI scan can be inaccessible for many.

The low maximum depth of the decision tree model likely affected it's performance, so that the model couldn't fully make us of the available features. The relatively small size of the dataset is also a limiting factor. More data points of patients not included in the dataset would assist in training a better model.

Code

The Python code (in a Jupyter notebook) and the dataset can be found in my public GitHub repository at `https://github.com/hakav/dementia-detection`.

References

[1] https://www.who.int/en/news-room/fact-sheets/detail/dementia

[2] Battineni, Gopi; Amenta, Francesco; Chintalapudi, Nalini (2019), "Data for: MACHINE LEARNING IN MEDICINE: CLASSIFICATION AND PREDICTION OF DEMENTIA BY SUPPORT VECTOR MACHINES (SVM)", Mendeley Data, V1, doi: 10.17632/tsy6rbc5d4.1

[3] Gopi Battineni, Nalini Chintalapudi, Francesco Amenta, Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM), Informatics in Medicine Unlocked, Volume 16, 2019, 100200, ISSN 2352-9148

[4] https://www.sciencedirect.com/topics/medicine-and-dentistry/clinical-dementia-rating

[5] https://www.mdpi.com/2076-3425/9/9/212/htm

[6] A. Jung, "Machine Learning: The Basics.", Springer, Singapore, 2022, https://alexjungaalto.github.io/MLBasicsBook.pdf

[7] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.