# Max Planck School of cognition: Artificial Intelligence and Intelligent Systems - end of course assessment

Anisbel Leon[1] and Also Hanna Wiedenhaus[1]

[1]Leibniz-Institut für Troposphärenforschung e.V., Permoserstraße 15, 04318 Leipzig

September 1, 2021

## Contents

# 1 Is it possible to estimate the state of dementia of a patient given the deficiency of vitamin B12 and the tendency to alcoholism?

## 1.1 Introduction

Dementia is a combination of symptoms of progressive decline in cognitive, emotional and social skills. One of the key symptoms is memory impairment. A few forms of dementia are reversible; for some other forms, limited therapeutic interventions are possible, but these can only delay the onset of certain symptoms. This report examines the proximate influences of lifestyle on the incidence of dementia. Metabolic B12 deficiency is widely prevalent and occurs in 10-40% of the population (Spence, 2016). It has long been associated with a variety of hematologic, neurologic, and psychiatric disorders. The described psychological effects associated with cobalamin deficiency range from mood swings to psychotic episodes to cognitive impairment (slow thinking, memory impairment, confusion, dementia) (Goebels and Soyka, 2000). The psychiatric syndrome most commonly associated with vitamin B12 deficiency in the medical literature is dementia (Cunha et al., 1995). Since this is one of the few treatable causes of dementia, extensive research is needed.

On the other hand, alcoholims is also a factor that could produce dementia. As it is often related to personality and behavioural changes, social and personal neglect, confabulation, lack of insight, empathy and emotional control seems to be a plausible main cause of the alcoholic dementia (Brun and Andersson, 2001).

Investigation of the relationship between vitamin B12 deficiency, alcoholism and dementia should be based on a broad data set. Therefore, the OASIS-3 (Open Access Series of Imaging Studies) dataset is used in this study. The dataset includes MRI and PET imaging and associated clinical data for 1098 participants. The study examined 605 cognitively normal adults and 493 individuals in various stages of cognitive decline ranging in age from 42 to 95 years over a 15-year period (LaMontagne et al., 2019).

## 1.2 Methods

### 1.2.1 Data overview

OASIS-3 Dataset

In the dataset, dementia status was assessed using the Clinical Dementia Rating Scale (CDR). Where CDR 0 represents normal cognitive function, CDR 0.5 for very mild impairment, CDR 1 for mild impairment, and CDR 2 for moderate dementia.Once a

participant reached CDR 2, he or she was no longer eligible for face-to-face assessment (LaMontagne et al., 2019). Data were collected according to the Uniform Data Set (UDS), which defines a common set of clinical observations collected longitudinally from ADC participants using standard methods (Morris et al., 2006). During the assessment of the participants, a diagnostic recording and an Interview was conducted. The results were recorded in a coded dementia diagnosis: in the OASIS "ADRC Clinical Data" dx1-dx5 data type. Diagnoses for this variable included "cognitively normal," "AD dementia," "vascular dementia," and contributing factors such as vitamin deficiency, alcoholism, and mood disorders. Diagnosis identification for variables dx1-dx5 is performed separately from UDS assessments (LaMontagne et al., 2019). For data preprocessing, participants with missing data were removed from the dataset. Furthermore, the occurrence of the diagnoses "Active B-12 Deficiency", "Remote B-12 Deficiency", "Active Alcoholism" and "Remote Alcoholism" were enumerated in the data set. To investigate the possible associations between dementia, vitamin B12 deficiency and alcoholism in the present dataset, we apply different machine learning approaches.

### 1.2.2   Data preparation

The variables we used from the dataset can be visualized in the pairplot below (Fig 1.). As we expected the variables that show more evident normal distribution are the "days since entry" and the "age at entry". As "cdr" depends on 4 criteria ctegories, the curve in the diagonal of the figure, shows that the CDR = 0 is more frequent in the pacients. On the other hand, the B12 deficiency and alcoholism tendencies are binary variables reporting less people with this problems.

```
<seaborn.axisgrid.PairGrid at 0x7f3e108a8a00>
```
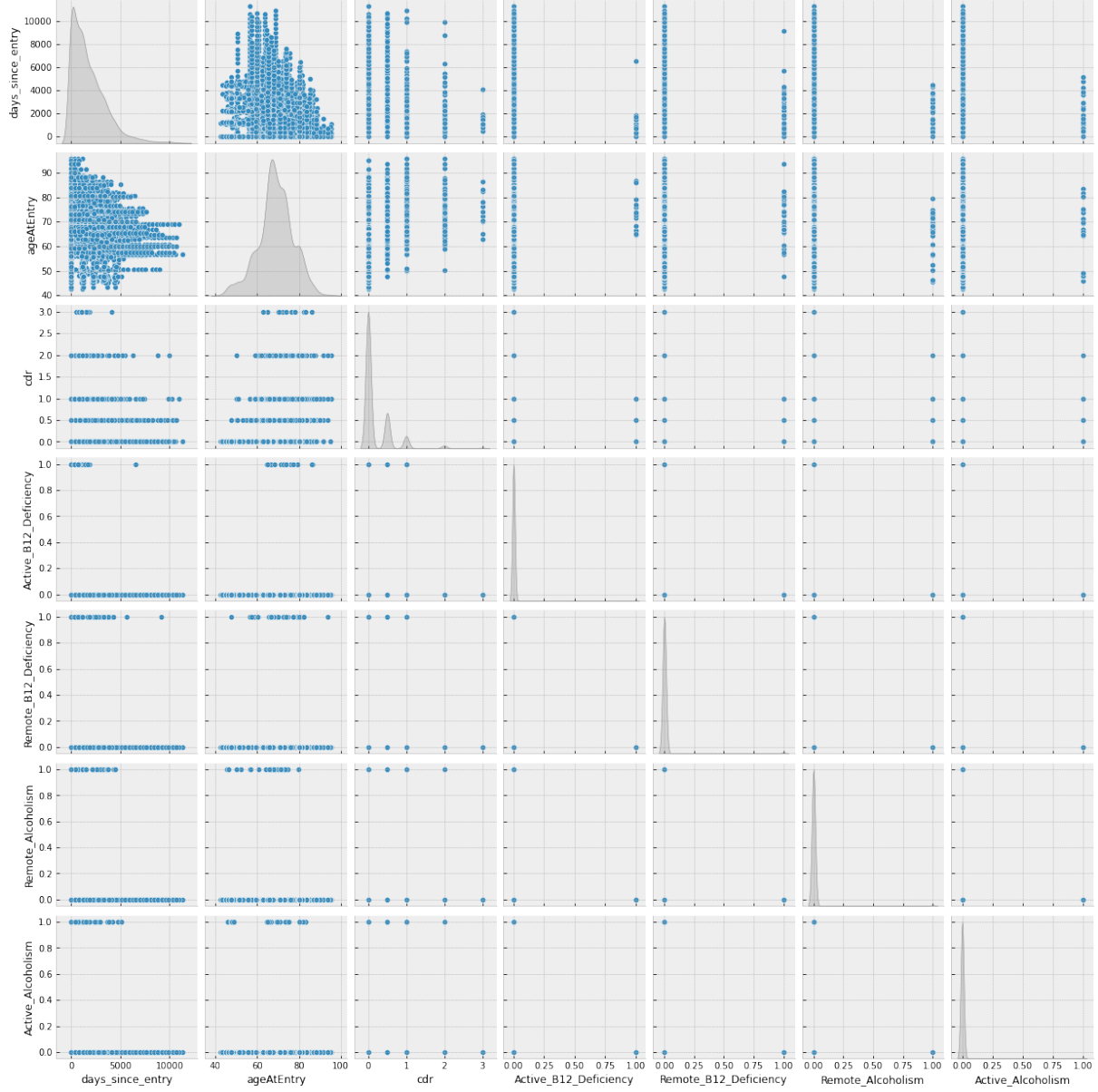
Fig 1. Plot of the data used in the present work.

As the distance metrics help algorithms to recognize and study a relationship between each of the elements of the dataset, we calculated some of them. A good distance metric could significantly help in improving the performance of Classification Machine Learning process.

For an easier analisis of the results, the data was normalized before calculating the metrics.

Three distance metrics were calculated: Cosine, Euclidean and Manhattan or City Blockn (Fig. 2).

The most common distance measure, Euclidean distance, is quite good and representative in our study of low dimention. On the other hand, since Manhattan distance does not calculate the shortest path possible, it is more likely to get higher values compared to the Euclidean distance. Although, Cosine distance shows similarities in the data and seems to be reasonable, we selected the first one (Euclidean) for further applications in the machine learning algorithms we used in this project.
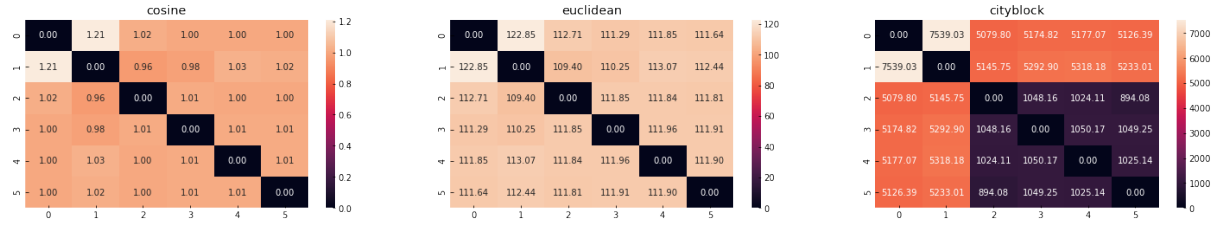


Fig 2. Distance metrics

There are a lot of Machine Learning methods to be applied based on supervised and unsupervised learning. In the present work we selected just two supervised learning methods: K-Nearest Neighbor (KNN) and Logistic Regression.

Both are clasification models and, despite their simplicity, they are very popular machine learning algorithms. KNN has proven to be incredibly effective at certain tasks in machine learning.

Our main goal from now on, is to find the best parameters to train KNN using some training set and get the highest performance when it is tested with the testing set.

Before we move forward, it is necessary to understand how these methods work.

1. The KNN algorithm uses 'feature similarity' to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set ( more ).

2. Logistic Regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems ( more ).

For KNN, the distance between the new point and each training point is calculated, in our case we selected Euclidean distance for this calculation ( more about the python method used ).

## 1.3   Results and Discussion

For validation, we used k-fold cross-validation. In this procedure, the data set is divided into training and test data. The technique has a single parameter called k, which refers to the number of groups into which the dataset is divided. We chose k=15, which is a 15-fold cross-validation. This means that each sample is included once in a test data set and is used k-1 times to train the model. Cross-validation is primarily used in applied machine learning to estimate the capabilities of a machine learning model. That is, a limited sample is used to estimate how the model will perform in general when used to make predictions on data that was not used during the training of the model.

Two analysis were made applying Cross-validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set:

- Train test split approach

- Kfold

Fig 3. represents the training and testing accuracy for different numbers of neighbors, K (from 0 to 14).

As it is expected the training score is allways higher than the testing score.The last one present a relatively low accuracy for k < 4 when it is compared with the training score. For K > 4 the performance seems to be nearly the same for any K.

This analysis provides valuable information, the highest testing accuracy occurs when K = 6 and for K > 4 is always bigger than 91%.

```
The number of neighbours in the KN Classifier with the highest testing accuracy
(91.542%) is:
 6
```
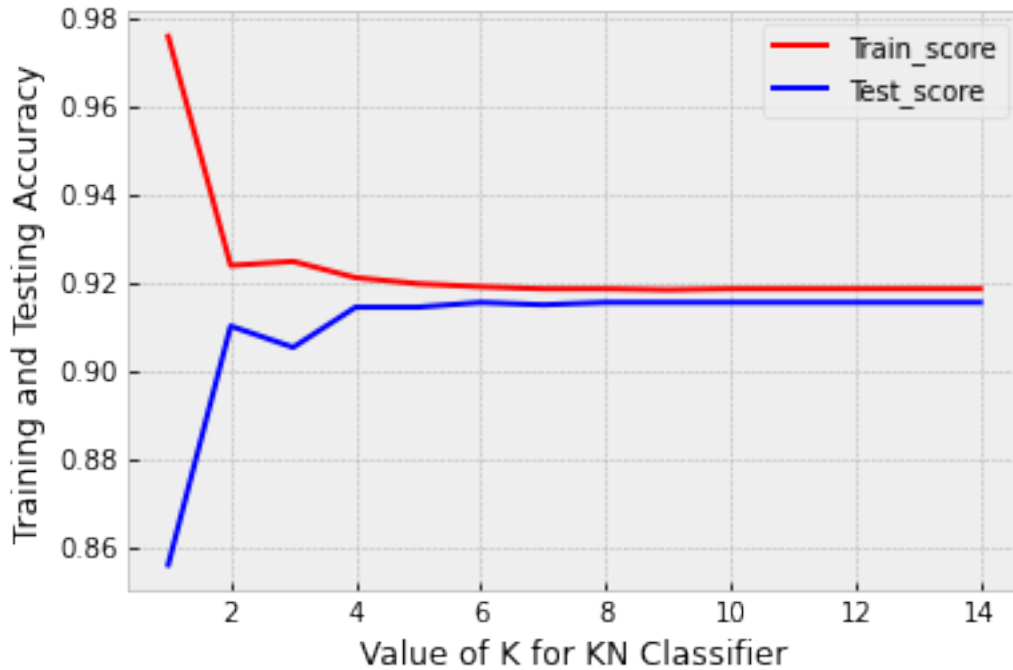
Fig 3. Accuracy values for each number of neighbours in the KN Classifier (from 0 to 14).

**However, we are probably not using a significant portion of our dataset. For more reliable results, we use kfold, to divide a data set into training and test sets.**

**Fig 4., shows in the upper pannel, a bar plot showing the performance of KNN for different K values. The error bars are quite low and in the table are presented the specific accuracy values and the standard deviation.**

**The lower left panel plot looks similar to the test_score curve from the previous analysis nevertheless, in this case the number of neighbours in the KN Classifier with the highest accuracy (91.758 % which is also a better performance than the best score form the previous analysis) is K = 8.**

```
The number of neighbours in the KN Classifier with the highest accuracy
(91.758%) is:
 8
```

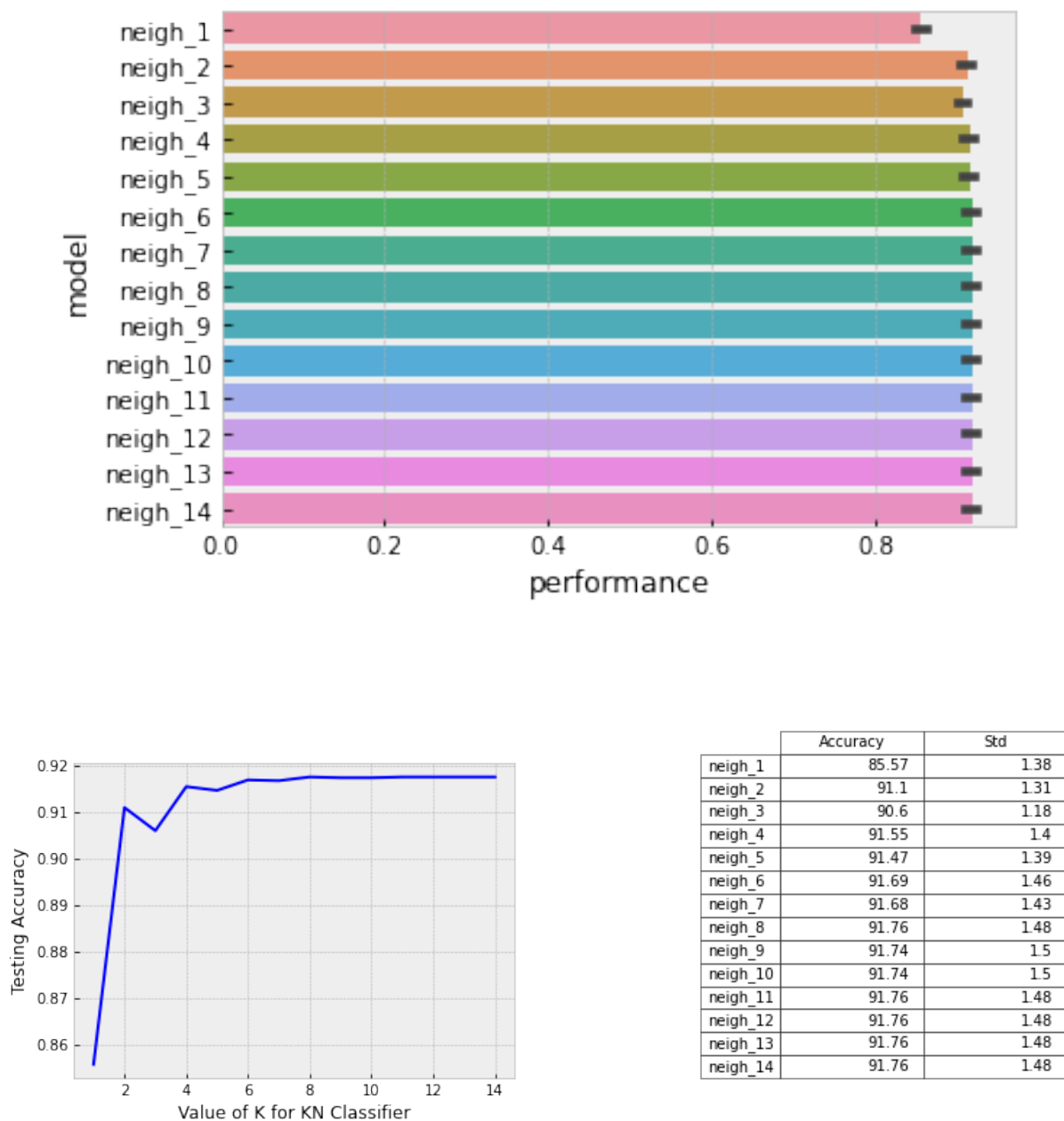| | Accuracy | Std |
|---|---|---|
| neigh_1 | 85.57 | 1.38 |
| neigh_2 | 91.1 | 1.31 |
| neigh_3 | 90.6 | 1.18 |
| neigh_4 | 91.55 | 1.4 |
| neigh_5 | 91.47 | 1.39 |
| neigh_6 | 91.69 | 1.46 |
| neigh_7 | 91.68 | 1.43 |
| neigh_8 | 91.76 | 1.48 |
| neigh_9 | 91.74 | 1.5 |
| neigh_10 | 91.74 | 1.5 |
| neigh_11 | 91.76 | 1.48 |
| neigh_12 | 91.76 | 1.48 |
| neigh_13 | 91.76 | 1.48 |
| neigh_14 | 91.76 | 1.48 |

Fig 4. Performance (upper pannel), testing accuracy curve (lower left panel) and Accuracy values and standar deviation (table) of KNN for different K values (0-14).

**So far we have a machine learning method that accurately represents and describes our data.**

**What about testing another method?**

As mentioned before, Logistic Regression is a very simple method which has two possible outcomes and uses a logistic function to model a binary dependent variable. We accomodate our the settings selecting the best solver for small datasets ( more about the python method used ).

Interestingly, applying this algorithm, we obtained the same precision as for the best KNN. This could be related to the simplicity of the data used.

The figure (Fig 5.) shows that both methods are good for estimation. The execution time was also measured, but in both cases it turned out to be very similar.

As the errors around the accuracy were low we considered not necessary to aply any other criteria.

```
KNeighborsClassifier(n_neighbors=8)
 Accuracy: 91.758% (1.485%)

LogisticRegression(solver='liblinear')
 Accuracy: 91.758% (1.485%)
```
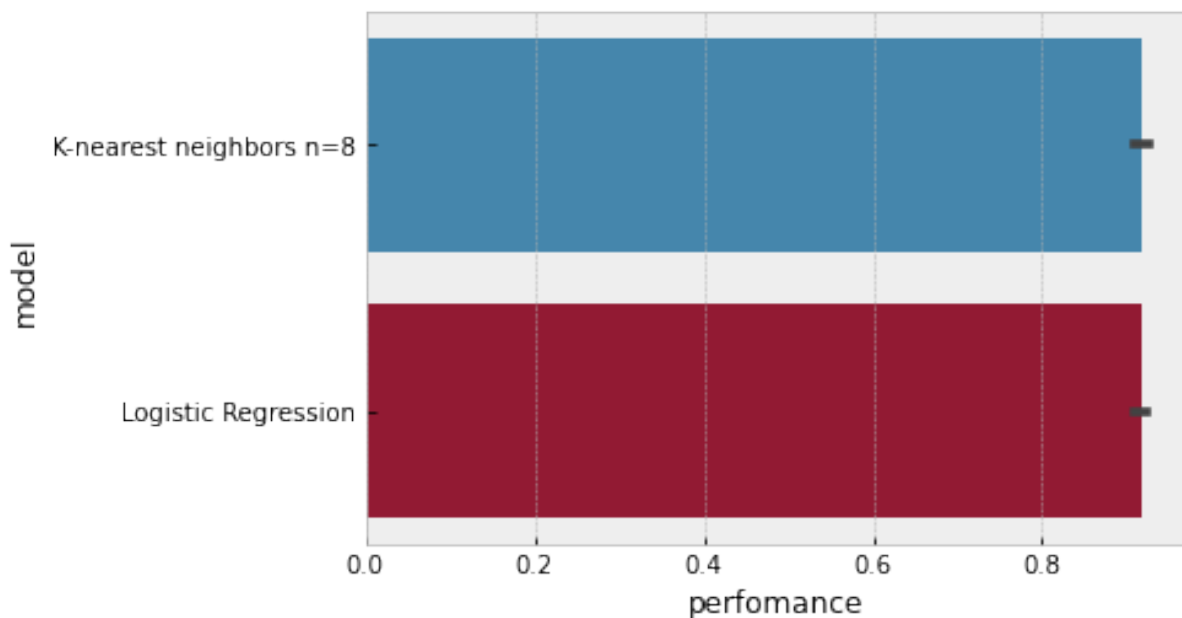


Fig 5. Error bars showing the performance of KNN for K = 8 and Logistic Regression.

## 1.4   Conclusions

In this work we aimed to investigate whether is possible to determine dementia tendencies in a patient of a certain age and vitamin B12 deficiency and alcoholism tendency information. We applied two machine learning approaches. One of them (K Neighbors Classifier (KNN)) was tested for various paramaters and was adjusted to get a better accuracy in the estimation. Our data was split into test and training sets and, both supervised clasiffication algorithms (KNN and Logistic Regression ) used presented high accuracy 91.758% and low errors indicating that could be applied for the estimation of Clinical Dementia Rating Scale (CDR).