CS 5710
Machine Learning

Spring 2023

Project Part-2 (Proposal + Increment)

EARLY-STAGE ALZHEIMER'S DISEASE PREDICTION USING MACHINE LEARNING

Submitted by

Mahendra Kumar Reddy Narapureddy (700741313)
Shiny Sherly Katuru (700744314)
Praveena Goli (700743010)
Sushrithareddy Sangam (700742861)

2) Goal & Objectives: -

Motivation

- Human instincts and standard measures often disagree in the current situation. Solving this problem requires the use of computationally intensive, non-traditional, and innovative approaches such as machine learning.
- Predictive and personalized medications are made possible by the use of machine learning techniques in disease imaging and prediction. This drift helps doctors choose treatments and health economists do analyses, as well as improving the quality of life for patients.
- Radiologists may not notice other medical issues when reviewing medical reports. As a result, several factors and circumstances are considered. Finding knowledge gaps and potential business prospects in relation to ML frameworks and EHR-derived data is the aim of this study.

SIGNIFICANCE:

- In older persons, Alzheimer's disease (AD) is the most common cause of dementia. Machine learning is currently being used to research metabolic disorders like Alzheimer's and diabetes that affect a sizable part of the global population. Each year, their incidence rates are rising alarmingly.
- So, we are using Machine Learning Algorithms, to supply better results for the existing problem above.

OBJECTIVES:

- Implement the algorithms in the real problems by understanding the decision tree algorithm & support vector Machine.

• Using hybrid algorithms and combining supervised and unsupervised learning, as well as ML and deep learning techniques, may improve outcomes.

FEATURES:

No. of visits, MR delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF were being used in the dataset to get better results accordingly with our existing problem.

Increment:

Dataset:

M/F	Gender
Age	Age
EDUC	Education in years
SES	Socioeconomic Status
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor
Group	Group
Hand	Right or left
Subject ID	Subject Identification
MRI ID	MRI Identification

Features in detail:

1. MRI ID:
   The Open Access Series of Imaging Studies (OASIS) project produced MRI-related data. using OASIS-1 cross-sectional and OASIS-2 longitudinal MRI datasets openly accessible data for instruction and testing with various machine learning models. Subjects in the longitudinal datasetFor each subject, including both men (n = 62) and women (n = 88), were right-handed, at least 3 or 4 bought T1-weighted brain MRI images bought during one of her MRI scans meetings. Each subject was scanned for a total of 373 imaging sessions in the longitudinal dataset.2 or more separate instances at least 1 year apart with an average lag of 719 days (default delay period is between 183 and 1707 days) until next visit.
2. Group:
   There are three groups in the dataset about the stages of Alzheimer's disease. It helps the model to predict which group is the given input data belongs.
   - Demented

- Non demented
- Converted

3. M/F:
   Gender of the patient. women seem more likely to develop AD than men. The reasons for this are still unclear.
4. Age:
   Age is the greatest risk factor after genetics and family history. The percentage of people who have Alzheimer's disease significantly increases with age. According to the prevalence section on page 19 (5%) of persons 65 to 74, 13.1% of adults 75 to 84, and 33.3% of adults 85 and older have Alzheimer's dementia. As the baby-boom generation ages, the number of Americans with Alzheimer's disease will significantly increase. It is important to keep in mind that Alzheimer's dementia is not a normal part of aging, and that aging does not cause the disease by itself.
5. EDUC:
   Years of education. Patient education helps manage chronic disease by informing and engaging patients in the treatment regimens and lifestyle changes needed to keep adverse outcomes at bay.
6. SES:
   Even with a high hereditary risk, people with high socioeconomic deprivation—as decided by income/wealth, unemployment rates, car/home ownership, and crowded housing—have a much higher risk of developing dementia than people with lower socioeconomic status. The part that socioeconomic status plays in dementia risk; and second, to enhance the body of evidence by looking into the link between socioeconomic characteristics and dementia-related deaths.
7. MMSE:
   A 30-point screening is the MMSE.Orientation, concentration, attention, verbal learning, naming, and VisuoConstruction (VC) were the six areas of measurement. MMSE examination for cognitive testing, nevertheless, can easily lead to the ceiling effect can be influenced by factors such as age, educational level, language, and cultural background. The 30-point [17-21] scale used to produce an evaluation or diagnostic is as follows:

   Score: 20 to 24 show mild dementia.
   • A score of 13 to 20 shows mild dementia.
   • Severe dementia is showed by a score of fewer than 12.
   Both sensitivity (71-92%) and specificity (56-96%) are high for the MMSE. It has good specificity (56–96%) and sensitivity (71–92%). one of the crucial elements
   MMSE scores are influenced by education level. Higher education recipients scored highly on the MMSE.

8. CDR:

   The Clinical Dementia Rating is a scale used to assess the severity of dementia. It is used to diagnose dementia, mainly Alzheimer's disease. CDR is expressed in 5 stages, CDR=0 is
   People without cognitive impairment, and four other items are:
   • 0 = Normal • 0.5 = very mild dementia • 1 = mild • 2 = moderate • 3 = Severe
9. eTIV:

Unlike brain volume, which should be vulnerable to neurodegeneration and atrophy with age, intracranial size measures give an estimate of the maximal size of the brain before disease. Understanding the pathophysiological characteristics of this disease requires evidence that the age at which symptoms first appear and intracranial size are related. Total intracranial volume (TIV) may also need to be taken into account as a potential confounding factor in future studies on disease prevention and epidemiological studies of AD risk factors if the link is proven to be statistically significant. Additionally, TIV is often used in studies of both global and focal atrophy in AD to account for head size on the grounds that it is unaffected by the presence or absence of the illness.5 In order to ensure that the assumptions are accurate, care must be used.

10. nWBV:

    As a percentage of all the voxels in the atlas-masked image that are appointed as grey or white matter, the normalized whole brain volume is expressed using an automated tissue segmentation procedure.
11. ASF:

    Scaling factor for Atlas. It is a calculated scaling factor that converts the atlas target from the native-space brain and skull.

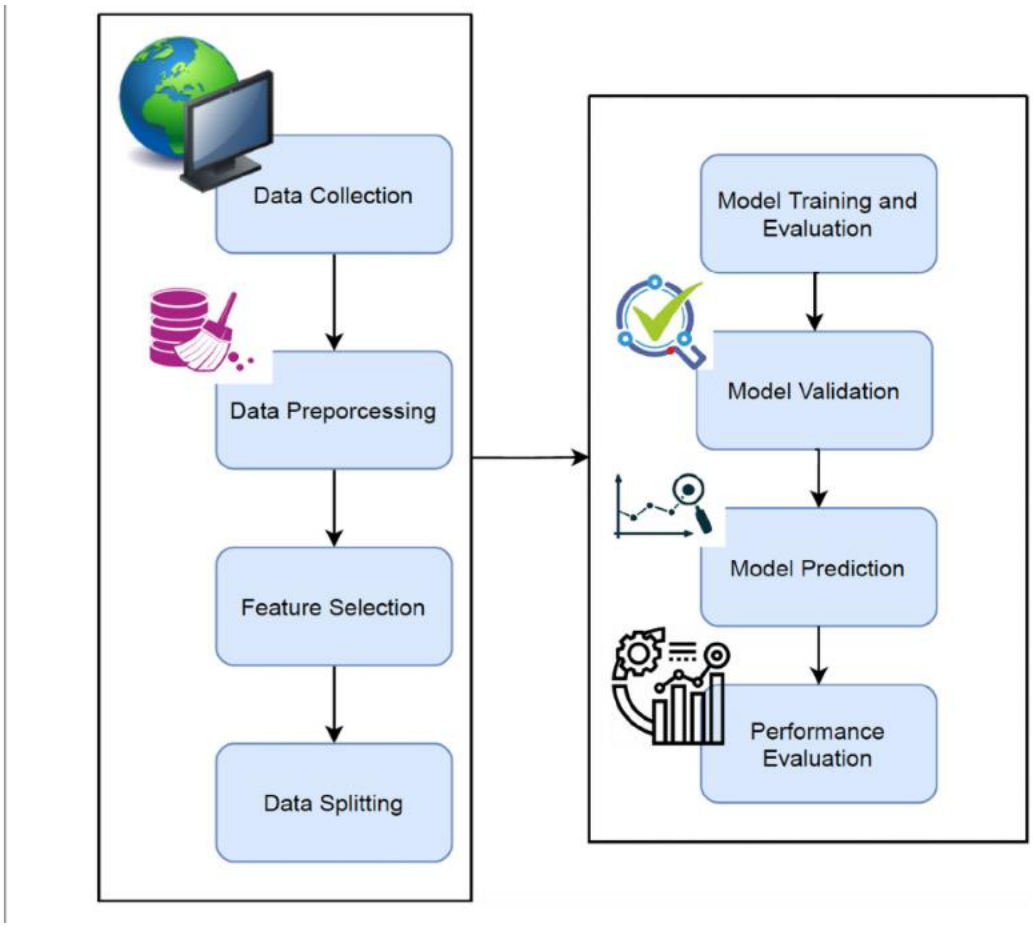
Analysis:

The connections between each characteristic have been discussed. The topics of MRIs and dementia are covered in this section. Before drawing or analyzing the correlation coefficients, we used the exploratory data analysis procedure 36, 37 to estimate them in order to establish a clear relationship between the data using graphs. That information may have been used later to help decide which approach to use to analyze it and to interpret its meaning. The maximum, mean, and median values are shown in the table

below.

Min	Max	Mean	Median
EDUC	7	22	14.2
SES	2	6	2.3
MMSE	16	30	26.2
CDR	0	1	0.3
eTIV	1,120	1,990	1,450
nWBV	0.55	0.81	0.7
ASF	0.87	1.43	1.3

IMPLEMENTATION:



Data preprocessing:

In this stage, the data was cleaned and preprocessed using a variety of data-mining techniques. In this context, missing values are processed, features are retrieved, altered, and so on. We found 9 entries in the SES column with missing data (34, 35). There are two approaches to this problem. Dropping the rows with empty values is the simplest method. Imputation (21), which entails substituting the missing values with their corresponding values, is the alternative method for filling in the gaps. We only have 140 measurements, so if we impute, the model ought to perform better. In the SES property, the 9 rows with missing values are deleted, and the median value is employed for imputation.

Feature selection:

The choice of features is crucial in machine learning. In this study, the clinical data are subjected to feature selection. where we have thousands of samples, is Alzheimer's disease. Three techniques are available for feature selection (22): filter methods, wrapper methods, and embedding methods. The filter method is a typical technique employed during the pre-processing step. Another technique that cores the feature subset is the use of wrapper methods. The filter and wrapper techniques are combined in the embedded method. The correlation coefficient, Information gain, and Chi-Square are the approaches of feature selection that have been used most often and successfully in this work.

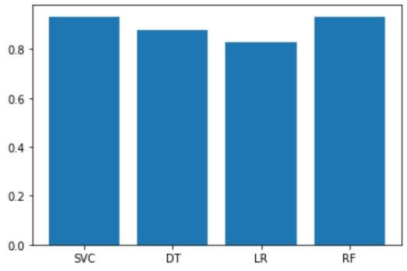
Preliminary Results:

As we have worked on four different algorithms, with the insights we have concluded by comparing the accuracies. To take the game further, we decided on doing a website to get the input result.

```
In [13]: print(results)
         {'SVC': 0.9333333333333333, 'DT': 0.88, 'LR': 0.8266666666666667, 'RF': 0.9333333333333333}

In [14]: import matplotlib.pyplot as plt
         names = list(results.keys())
         values = list(results.values())

         plt.bar(range(len(results)), values, tick_label=names)
         plt.show()
```



Work completed:

Description:

Decision Tree: This model splits the data according to the feature's cutoff values and is based on a tree structure. Instances are split up to form subsets. The term "leaf node" refers to a leaf. Internal nodes are the word for intermediate subsets. A decision tree is most useful when there is strong interaction between the features and the target.

```
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
y_pred = dt.predict(X_test)
accuracy=metrics.accuracy_score(y_test, y_pred)
results.update({"DT":accuracy})
print("Accuracy:",accuracy)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.8666666666666667
[[32  0  3]
 [ 1 32  3]
 [ 1  2  1]]
              precision    recall  f1-score   support

         0.0       0.94      0.91      0.93        35
         1.0       0.94      0.89      0.91        36
         2.0       0.14      0.25      0.18         4

    accuracy                           0.87        75
   macro avg       0.68      0.68      0.67        75
weighted avg       0.90      0.87      0.88        75
```

(RF) Random Forest A random forest model performs better than a decision tree model since it is not overfit. Different decision tree types that are only marginally distinct from one another make up random forest-based models. Using a majority vote process, the ensemble produces forecasts based on each unique decision tree model (bagging). As a result, each tree's correctness

is preserved while less overfitting takes place.

```
rclf = RandomForestClassifier(n_estimators=100,n_jobs=-1)
rclf.fit(X_train, y_train)
y_pred = rclf.predict(X_test)
accuracy=metrics.accuracy_score(y_test, y_pred)
results.update({"RF":accuracy_score(y_test,y_pred)})
print("Accuracy:",accuracy)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.9466666666666667
[[35  0  0]
 [ 0 35  1]
 [ 1  2  1]]
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99        35
         1.0       0.95      0.97      0.96        36
         2.0       0.50      0.25      0.33         4

    accuracy                           0.95        75
   macro avg       0.81      0.74      0.76        75
weighted avg       0.93      0.95      0.94        75
```

SVM, short for support vector machine The data points are categorized using the appropriate hyper planes in a multidimensional space using this method. With the use of SVM (25), we attempt to identify a hyperplane that divides instances of two categories of variables that are found in adjacent vector clusters, one on each side. The vectors closest to the hyperplane are known as support vectors. In SVM, training and test sets of data are both utilized. By target values and attributes, training data categories are divided. Using test data as input, a model is created

using SVM to predict target values.

```
svm = SVC(kernel = 'linear', C = 1, gamma = 1)
svm.fit(X_train, y_train)
y_pred = svm.predict(X_test)
accuracy=metrics.accuracy_score(y_test, y_pred)
results.update({"SVC":accuracy})
print("Accuracy:",accuracy)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.9333333333333333
[[35  0  0]
 [ 0 34  2]
 [ 1  2  1]]
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99        35
         1.0       0.94      0.94      0.94        36
         2.0       0.33      0.25      0.29         4

    accuracy                           0.93        75
   macro avg       0.75      0.73      0.74        75
weighted avg       0.92      0.93      0.93        75
```

Logistic Regression

The logistic model, often known as the logit model, is a statistical model that predicts an event's likelihood by making the event's log-odds a linear combination of one or more independent variables. Logistic regression, often known as logit regression, is a technique used in regression analysis to estimate a logistic model's parameters (the coefficients in the linear combination).

```
lr = LogisticRegression()
lr.fit(X_train, y_train)
y_pred = lr.predict(X_test)
accuracy=metrics.accuracy_score(y_test, y_pred)
results.update({"LR":accuracy_score(y_test,y_pred)})
print("Accuracy:",accuracy)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.8266666666666667
[[33  2  0]
 [ 4 29  3]
 [ 3  1  0]]
              precision    recall  f1-score   support

         0.0       0.82      0.94      0.88        35
         1.0       0.91      0.81      0.85        36
         2.0       0.00      0.00      0.00         4

    accuracy                           0.83        75
   macro avg       0.58      0.58      0.58        75
weighted avg       0.82      0.83      0.82        75
```

Responsibility (Task, Person)

Roles and responsibilities:

Mahendra Reddy: Data collection and worked on Random Forest.

Shiny sherly: exploratory data analysis and worked on decision tree.

Praveena goli: data visualization and worked on support vector machine.

Sushritha: comparing the accuracies and worked on logistic regression.

Contributions (members/percentage)

Mahendra Kumar Reddy: 25%

Shiny Sherly: 30%

Praveena goli: 25%

Sushmita: 20%

Work to be completed:

As we have already completed 80% of our project, we are still planning on adding two more algorithms and a website to get the output result for the given input. We are planning on adding GaussianNaivebayes, and K-nearest neighbors, adaboost, XG boost.

Roles and responsibilities:

Mahendra Kumar Reddy: K- nearest neighbors and working on the website.

Shiny sherly katuru: GaussianNaivebayes and working on the website.

Praveen goli: adaboost and working on the website.

Sushritha: XG boost and working on the website.

Issues/Concerns

1. Feature extraction
2. Comparing the accuracies
3. incorporating the website into the model

References/Bibliography

1. Sivakani GA, Ansari R. Machine learning framework for implementing Alzheimer's disease. Int Conferen Commun Signal Process. (2020)

2. Khan P, Kader MF, Islam SR, Rahman AB, Kamal MS, Toha MU, et al. Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances. IEEE Access. (2021) 9:37622– 55. doi: 10.1109/ACCESS.2021.3062484

3. Martinez-Murcia FJ, Ortiz A, Gorriz JM, Ramirez J, Castillo-Barnes D. Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. IEEE J Biomed Health Inform. (2020) 24:17–26. doi: 10.1109/JBHI.2019.2914970

4. Prajapati R, Khatri U, Kwon GR. "An efficient deep neural network binary classifier for alzheimer's disease classification," In: International Conference on Artificial Intelligence in Information and Communication (ICAIIC). (2021), p. 231–234.

5. Helaly HA, Badawy M, Haikal AY. Deep learning approach for early detection of Alzheimer's disease. Cogn Computing. (2021) 21:1–17. doi: 10.1007/s12559-021-09946-2

6. Yaffe K. Modifiable risk factors and prevention of dementia: what is the latest evidence. JAMA Intern Med. (2018) 178:281–

2. doi: 10.1001/jamainternmed.2017.7299

7. Livingston G, Sommerlad A, Orgeta V, Costafreda SG, Huntley, D, et al. Dementia prevention, intervention, and care. The Lancet. (2017) 390:2673– 73. doi: 10.1016/S0140-6736<17>31363-6

8. O'Donnell CA, Manera V, Köhler S, Irving K. Promoting modifiable risk factors for dementia: is there a role for general practice? British J General Pract. (2015) 65:567–8. doi: 10.3399/bjgp15X687241

9. Sulaiman N, Abdulsahib G, Khalaf O, Mohammed MN. "Effect of Using Different Propagations ofOLSR andDSDVRouting Protocols", In Proceedings of the IEEE International Conference on Intelligent Systems Structureing and Simulation. (2014), pp. 540-5.

10. Deckers K, van Boxtel MP, Schiepers OJ, de Vugt M, Muñoz Sánchez JL, Anstey KJ, et al. Target risk factors for dementia prevention: a systematic review and Delphi consensus study on the evidence from observational studies. Int J Geriatric Psychiatry. (2015) 30:234–46. doi: 10.1002/gps.4245

11. Schiepers OJ, Köhler, S., Deckers K, Irving K, O'donnell CA, Van den Akker, et al. Lifestyle for Brain Health (LIBRA): a new model for dementia prevention. Int J Geriatric Psychiatry. (2018) 33:167–75. doi: 10.1002/gps.4700

12. Vos SJ, Van Boxtel MP, Schiepers OJ, Deckers K, De Vugt M, Carrière I, et al. Modifiable risk factors for prevention of dementia in midlife, late life and the oldest-old: validation of the LIBRA Index. J Alzheimer's Dis. (2017) 58:537–47. doi: 10.3233/JAD-161208
13. Osamh Khalaf I, Ghaida M, Abdulsahib D. Energy efficient routing and reliable data transmission protocol in WSN. Int J Adv Soft Comput Applicat. (2020) 12:45–53.
14. National Academies of Sciences, Engineering, and Medicine. Preventing cognitive decline and dementia: A way forward. London: The National Academies Press (2018).