Statistical Learning and Data Mining

Challenge 1

Classification challenge on Alzheimer's Disease using MRIs and Gene Expression data

Taofik Ahmed Suleiman

Master Student in Medical Imaging and Applications

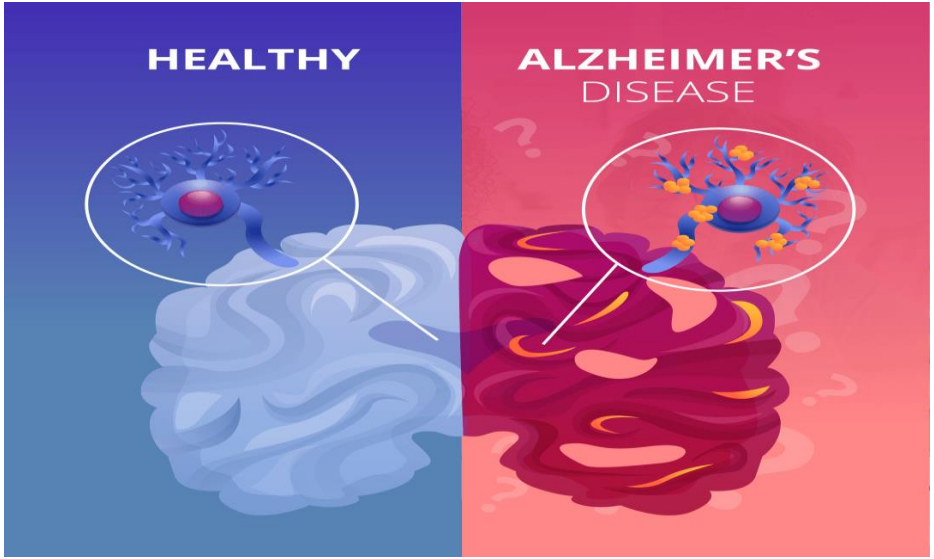
University of Cassino, Italy

Introduction

- Alzheimer's disease is a major public health concern affecting millions of people worldwide, consequently, early diagnosis is crucial for effective treatment.

- In this classification challenge, we used MRIs and gene expression data to classify patients into three macro-stages of Alzheimer's disease:

1. CTL (Controls): No deficit
2. MCI (Mild Cognitive Impairment): Few deficits
3. AD (Alzheimer's Disease): Dementia



Dataset Overview

For this challenge, a pair of three datasets were used. This consists of three train datasets with each corresponding to their equivalent test data.

- The summary of the datasets and their respective feature sizes are:
  1. ADCTLtrain = 430 features + 1 label

```
> str(ADCTL)
'data.frame':   164 obs. of  431 variables:
```

  2. ADMCItrain = 64 features + 1 label

```
> str(ADMCI)
'data.frame':   172 obs. of  65 variables:
```

  3. MCICTLtrain = 594 features + 1 label

```
> str(MCICTL)
'data.frame':   172 obs. of  595 variables:
```

- These features include data from demographic, clinical, CSF, medical imaging, and transcriptomics. With these large feature there is a need to finding an optimal solutions to three classification problems:
  1. AD vs CTL
  2. AD vs MCI
  3. MCI VS CTL

Methodology

The methodology applied for this research can be summarized below:

- Min-Max Normalization:
  - Due to the varying scales of each of the features in the dataset, min-max normalization was applied to ensure that each feature contributes equally during the model training.

- Principal Component Analysis (PCA):
  - The feature space for each classification problem (ADCTL, ADMCI, and MCICTL) was large, this can lead to potential overfitting of the model. To address this issue, PCA was applied to reduce the dimensionality of the feature space.

- Cross-Validation:
  - Firstly the training dataset was split in the ratio 70:30 to create a separate testing set, and then 5-fold cross-validation repeated 3 times is applied to the 70% training set to assess the performance of the model while avoiding information leakage. The process was repeated 3 times to obtain more reliable performance estimates.

- Logistic Regression:
  - Logistic regression was employed for each classification problem (ADCTL, ADMCI, and MCICTL). After trying several models such as KNN, SVM, and random forest, optimal performance was achieved with logistic regression for each cases.

Implementation and Results

- The implementation of the work was carried out in R. It involved splitting the training dataset into 30% for testing and 70% for training the model. To ensure robustness and generalize the performance of the model, 5-fold cross-validation techniques were employed during the training phase. The results obtained from the 5-fold cross-validation showed consistent and comparable accuracy across all folds, indicating the stability and reliability of the trained model.
- The trained and validated model was then used to predict the 30% test data, which had not been seen during the training phase. The goal was to evaluate the model's performance on unseen data, simulating real-world scenarios. The predictions generated for each classification problem were analyzed, and the following results were obtained:

1. **ADCTL**

Accuracy	Sensitivity	Specificity	Precision	F1_Score	AUC	MCC	Balanced_Accuracy
0.875	0.8333333	0.9166667	0.9090909	0.8695652	0.875	0.7526178	0.875

2. **ADMCI**

Accuracy	Sensitivity	Specificity	Precision	F1_Score	AUC	MCC	Balanced_Accuracy
0.7843137	0.75	0.8148148	0.7826087	0.7659574	0.7824074	0.5665662	0.7824074

3. **MCICTL**

Accuracy	Sensitivity	Specificity	Precision	F1_Score	AUC	MCC	Balanced_Accuracy
0.8627451	0.9259259	0.7916667	0.8333333	0.877193	0.8587963	0.7277717	0.8587963

Discussion and Conclusion

- The obtained results revealed promising performance across all three classification problems, with AUC ranging from 78.24% to 87.5% and MCC ranging from 56.66% to 75.26% on an unseen 30% split data from the training set. These results as evident in the various performance metrics used indicate the effectiveness of the models in classifying patients into their respective macro-stages of Alzheimer's Disease.

- In conclusion, this classification challenge demonstrated the application of MRIs and gene expression data for the early diagnosis of Alzheimer's Disease. Through the implementation of logistic regression models and preprocessing techniques such as min-max normalization and PCA, we were able to obtain a significantly stable performance across multiple folds of cross-validation. This model was also tested on unseen data set to further validate the model.