Statistical Learning and Data Mining

Final Assignment

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

Experimenting on different classifiers and feature selection methods for 3 different binary classification problems

Rachika Elhassna Hamadache

General Approach

In order to solve these classification tasks, we will explore different classifiers and rely on the following major steps : *Data Preprocessing*, *Training models*, *Evaluating* and *Choosing the Best Model*, and *Predicting*.

1. Data Preprocessing :

Prior to training our classifiers, and after making sure that the data has *no missing values* nor *categorical features* to take care of, we apply some *preprocessing* steps to prepare the data and improve its quality. These steps involves :

- Search for Outliers : We will focus on 2 methods to detect outliers : *'Grubbs Test'* and *'Rosner Test'*. As there's no strict rule whether those samples should be removed or not, we will compare the scores in both cases and opt for the most suitable one for each task.

- Feature Selection and Transformation : Different methods can be used to reduce the number of features in the data and make it more relevant. We will focus on the following ones :
  - **Highly Correlated Features :** By computing the *correlation matrix* of predictors and via the *Caret*'s built-in function *'findCorrelation'* with cutoff=0.9, we extract those features and remove them from the initial Train set.
  - **Recursive Feature Elimination :** The *'rfe'* function selects the *most relevant variables* in each model, based on *Cross-Validation* and on the *'AUC'* metric. The retained variables will be used afterward to train that model.
  - **Principal Component Analysis :** This algorithm is used to reduce the feature space dimensionality and produce uncorrelated features. We will use *PCA* during training through *Caret*'s built-in function *'train'* in the *'preProcess'* options.

→ Note : All features are *scaled* and *centered* during training.

General Approach

2. Models Training :

We will explore and train different classifiers : *LDA, GLM, KNN, NB, RF* and *SVM* (with *linear* and *radial* kernels), then compare their performance scores to decide which model is best fit for prediction. For each classifier, we will rely on :

- **Training through Cross-Validation :** Using the *Caret* built-in function *'train'*, and choosing the resampling method *'repeatedCv'* in the *'trainControl'* function, we train our models on 10 folds and obtain the final models.
- **Preprocessed Training data :** In order to evaluate the effect of preprocessing steps on each classifier, we train the models on the following transformed data (both *with* and *without outliers*) → Train set with *all features*, Train set with *RFE selected features*, Train set *without highly correlated features*, and Train set with *PCA*.

3. Models Evaluation :

Using the *mean* of the obtained results on each fold in each trained classifier, we compute some metrics in order to compare between each preprocessing method from each model and chose the most promising one for prediction (focusing mainly on *AUC* and *MCC scores*).

4. Predictions :

After choosing the best model for each classification problem, we predict the classes from the Test set using the function *'predict'* in R.

1. AD vs CTL Classification Problem - Results

- ## Preprocessing :

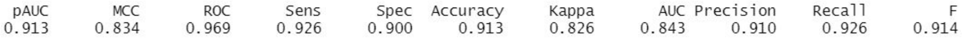
The preprocessing steps led to detect 46 and 13 outliers by Grubbs and Rosner's Tests respectively, and 112 highly correlated features *[fig1]*.

- ## Model Training :

After training each classifiers with the different preprocessing methods, we choose the best model from each classifier based on its *AUC* and *MCC* metrics and plot the results *[fig3]*.

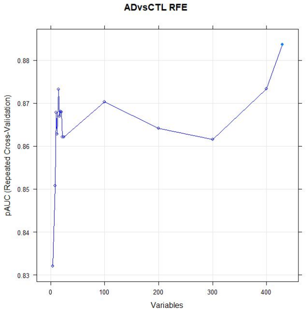
- ## Conclusion :

For this problem and for the available data, the *SVM* classifier with *linear* kernel appears to surpass other classifiers and have the best results *[fig4]* using *429 RFE-selected features* from the *initial* Train data *without removing outliers* *[fig2]*.

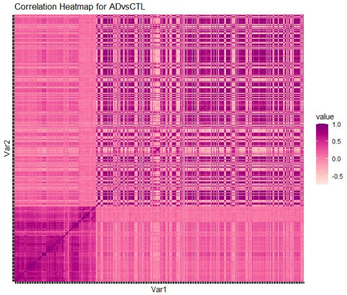


```
 pAUC       MCC       ROC      Sens      Spec  Accuracy     Kappa       AUC Precision    Recall         F 
0.913     0.834     0.969     0.926     0.900     0.913     0.826     0.843     0.910     0.926     0.914 
```

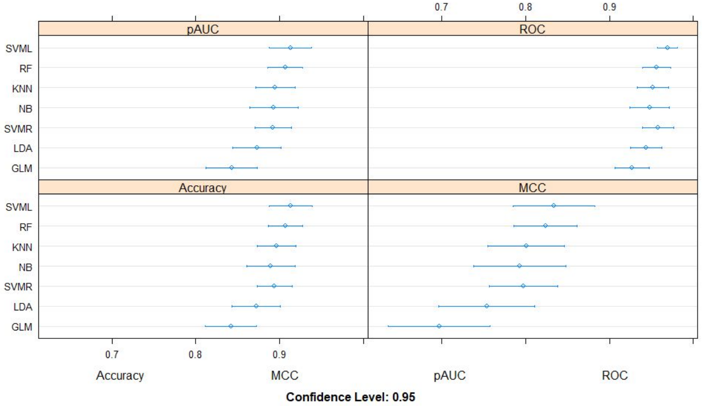
[fig4] - Scores from the chosen best model



[fig2] - Plot of AUC for different number of selected variables (RFE)



[fig1] - Correlation Matrix Heatmap



[fig3] - Dotplot of the best fitted models from each classifier

2. AD vs MCI Classification Problem - Results

- **Preprocessing :**

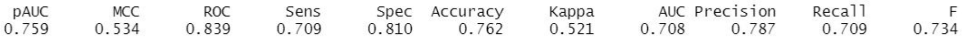
The preprocessing steps led to detect 7 and 8 outliers by Grubbs and Rosner's Tests respectively, and 10 highly correlated features *[fig5]*.

- **Model Training :**

After training each classifiers with the different preprocessing methods, we choose the best model from each classifier based on its *AUC* and *MCC* metrics and plot the results *[fig7]*.

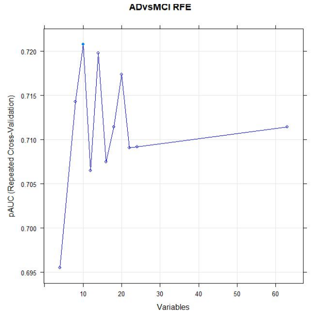
- **Conclusion :**

For this problem and for the available data, the *SVM* classifier with *radial* kernel appears to surpass other classifiers and have the best results *[fig8]* using *10 RFE-selected features* from *Rosner's outliers removed* Train data *[fig6]*.

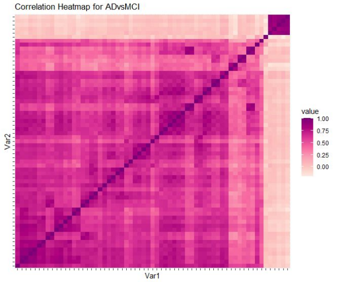


```
 pAUC       MCC       ROC      Sens      Spec  Accuracy     Kappa       AUC Precision    Recall         F
0.759     0.534     0.839     0.709     0.810     0.762     0.521     0.708     0.787     0.709     0.734
```

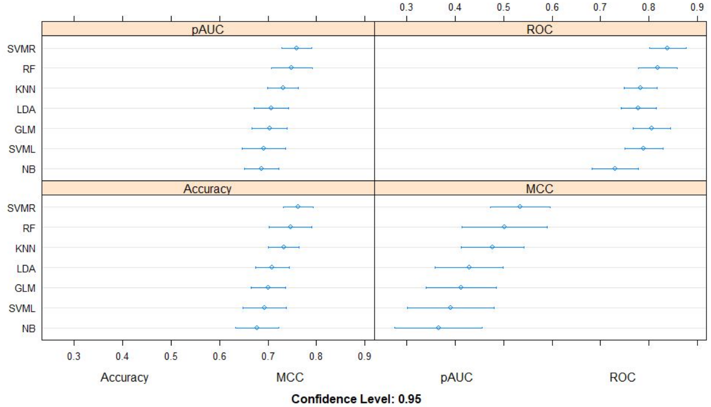
[fig8] - Scores from the chosen best model



[fig6] - Plot of AUC for different number of selected variables (RFE)



[fig5] - Correlation Matrix Heatmap



[fig7] - Dotplot of the best fitted models from each classifier

3. MCI vs CTL Classification Problem - Results

- ## Preprocessing :

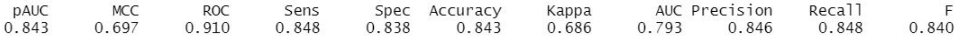
The preprocessing steps led to detect 83 and 12 outliers by Grubbs and Rosner's Tests respectively, and 210 highly correlated features *[fig9]*.

- ## Model Training :

After training each classifiers with the different preprocessing methods, we choose the best model from each classifier based on its *AUC* and *MCC* metrics and plot the results *[fig11]*.

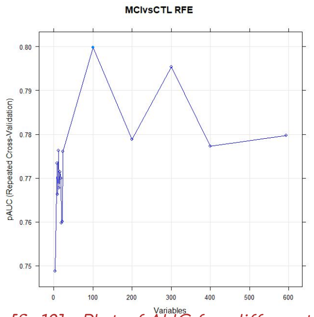
- ## Conclusion :

For this problem and for the available data, the *SVM* classifier with *radial* kernel appears to surpass other classifiers and have the best results *[fig12]* using *100 RFE-selected features* from *Rosner's outliers removed* Train data *[fig10]*.

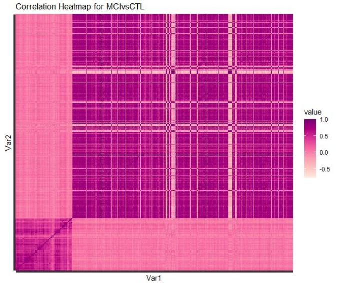


```
 pAUC       MCC       ROC      Sens      Spec  Accuracy     Kappa       AUC Precision    Recall         F
0.843     0.697     0.910     0.848     0.838     0.843     0.686     0.793     0.846     0.848     0.840
```

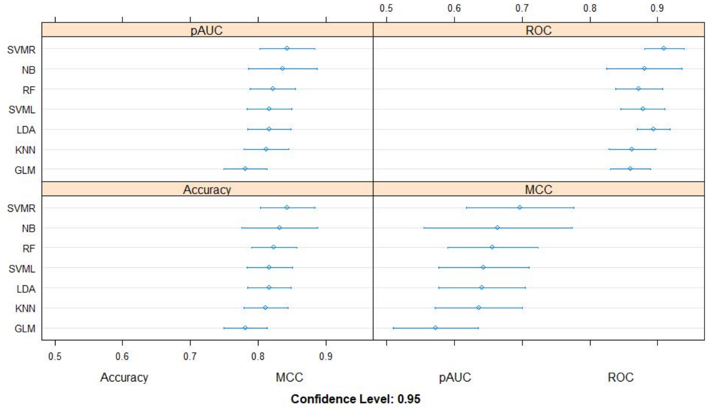
[fig12] - Scores from the chosen best model



[fig10] - Plot of AUC for different number of selected variables (RFE)



[fig9] - Correlation Matrix Heatmap



[fig11] - Dotplot of the best fitted models from each classifier