Classification Challenge:

Alzheimer’s Disease Classification Using MRIs and Gene Expression Data

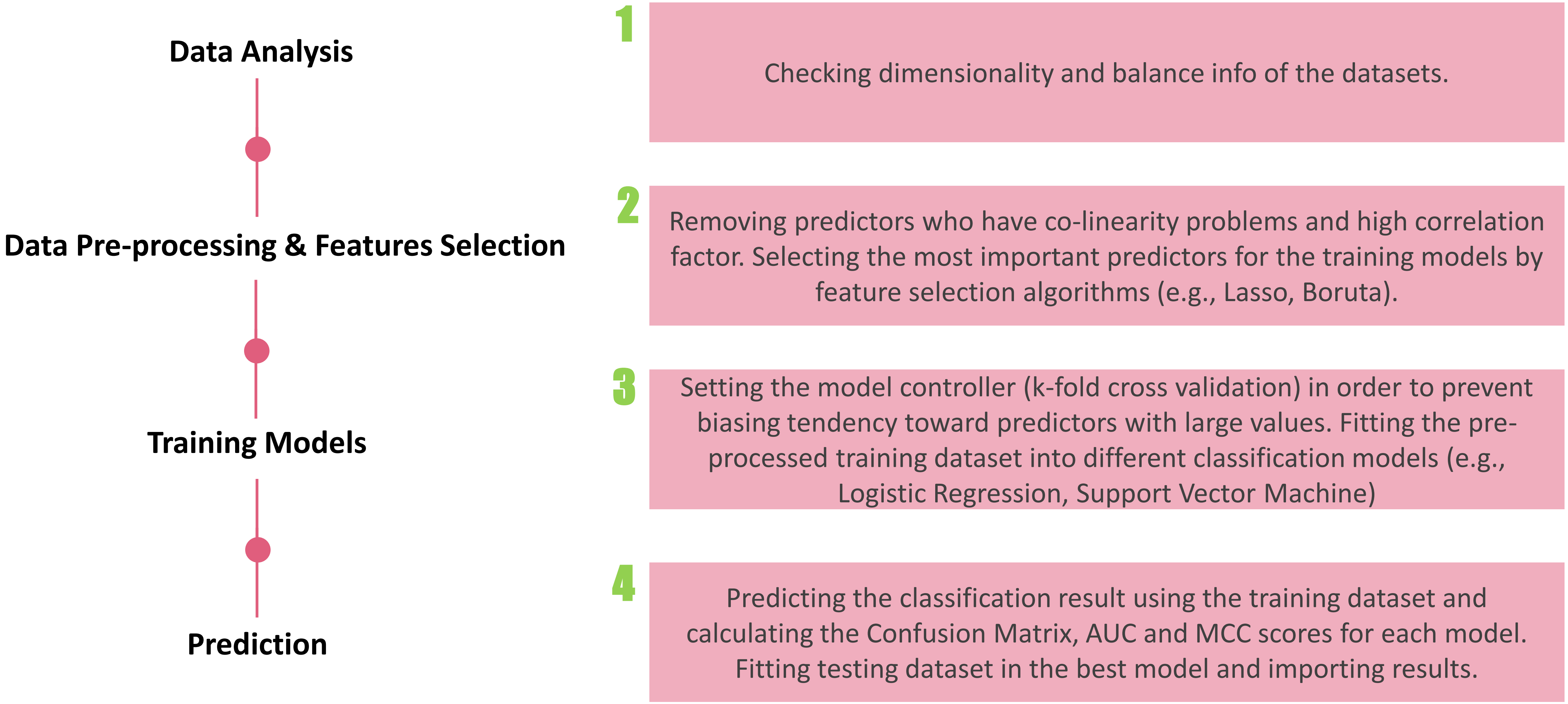
MD IMRAN HOSSAIN

Statistical Learning & Data Mining

Erasmus Mundus Joint Master in Medical Imaging and Applications

University of Cassino and Southern Lazio, Italy

Methodology



Data Analysis

Data Pre-processing & Features Selection

Training Models

Prediction

1 Checking dimensionality and balance info of the datasets.

2 Removing predictors who have co-linearity problems and high correlation factor. Selecting the most important predictors for the training models by feature selection algorithms (e.g., Lasso, Boruta).

3 Setting the model controller (k-fold cross validation) in order to prevent biasing tendency toward predictors with large values. Fitting the pre-processed training dataset into different classification models (e.g., Logistic Regression, Support Vector Machine)

4 Predicting the classification result using the training dataset and calculating the Confusion Matrix, AUC and MCC scores for each model. Fitting testing dataset in the best model and importing results.

Data Analysis

Dimensionality

Inspected number of predictors, **p** and number of samples, **n** in each dataset:

- ADCTL Dataset: **n** = 164; **p** = 429 → **very high** dimensionality **(p >> n)**
- ADMCI Dataset: **n** = 172; **p** = 63 → **low** dimensionality **(p < n)**
- MCICTL Dataset: **n** = 172; **p** = 593 → **very high** dimensionality **(p >> n)**

Data Pre-processing & Features Selection

Co-linearity

Firstly, Variance Inflation Factor (VIF) is used to identify predictors who have co-linearity problems. After that, predictors having co-linearity problems are eliminated from the training dataset. As a result, the number of predictors, **p** is reduced in the dataset. The number of predictors having co-linearity problem are:

- ADCTL Dataset: **110** predictors have co-linearity problems **(hence eliminated)**
- ADMCI Dataset: **8** predictors have co-linearity problems **(hence eliminated)**
- MCICTL Dataset: **192** predictors have co-linearity problems **(hence eliminated)**

Correlation

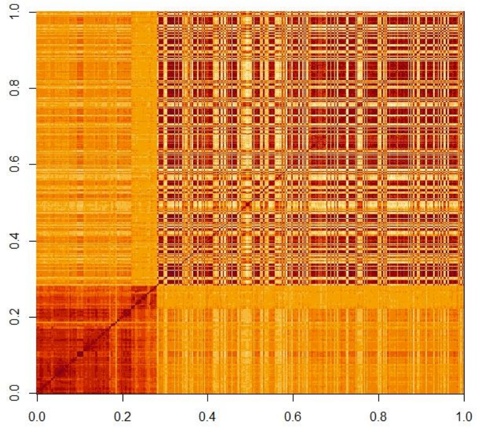
At first, the correlation matrix of predictors of each dataset is calculated and then removed the predictors who have more than 75% correlation factor from the dataset. In addition, the correlation matrices of datasets are plotted and noticed the presence of high correlation between some pairs of the feature variable (red dark) in all task.

- ADCTL Dataset: **184** predictors have correlation factor of **more than 0.75**
- ADMCI Dataset: **33** predictors have correlation factor of **more than 0.75**
- MCICTL Dataset: **315** predictors have correlation factor of **more than 0.75**

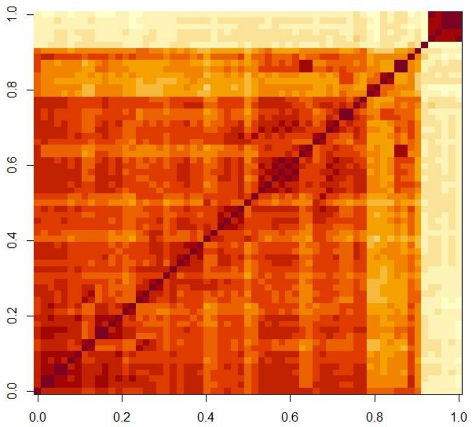
Balance

Observed the number of samples in each class:

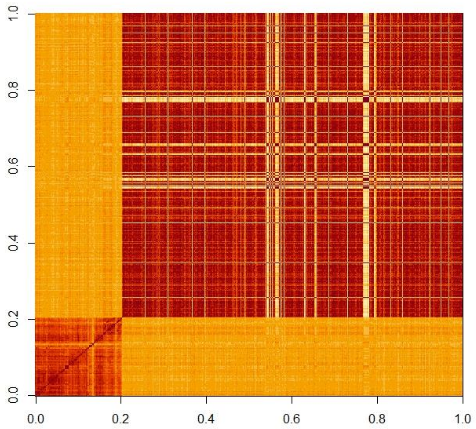
- ADCTL Dataset: **AD** = 81; **CTL** = 83 → **balanced** dataset
- ADMCI Dataset: **AD** = 82; **MCI** = 90 → **balanced** dataset
- MCICTL Dataset: **MCI** = 90; **CTL** = 82 → **balanced** dataset



Correlation Matrix of ADCTL Predictors



Correlation Matrix of ADMCI Predictors



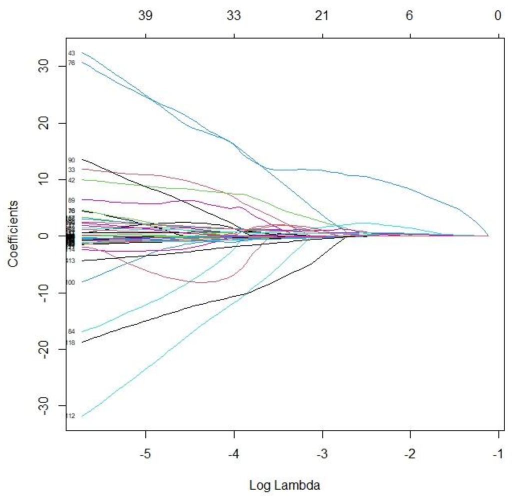
Correlation Matrix of MCICTL Predictors

Data Pre-processing & Feature Selection Methods

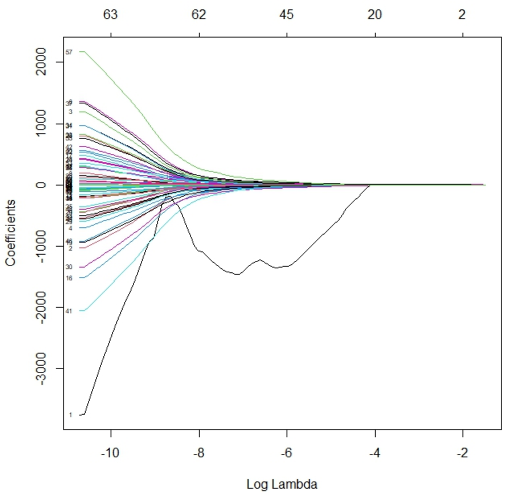
Lasso Regression

After removing features based on co-linearity and correlation of predictors, still overfitting is observed in some models due to a huge number of predictors, **p** compared to the total number of samples, **n**. Hence, Lasso regression is used in order to select less important predictors which are removed from the training dataset. The number of predictors after Lasso feature selection are:

- ADCTL Dataset: only **24** out of **429** predictors are available for training models
- ADMCI Dataset: only **9** out of **63** predictors are available for training models
- CTLMCI Dataset: only **6** out of **593** predictors are available for training models



Lasso Regression for ADCTL Dataset

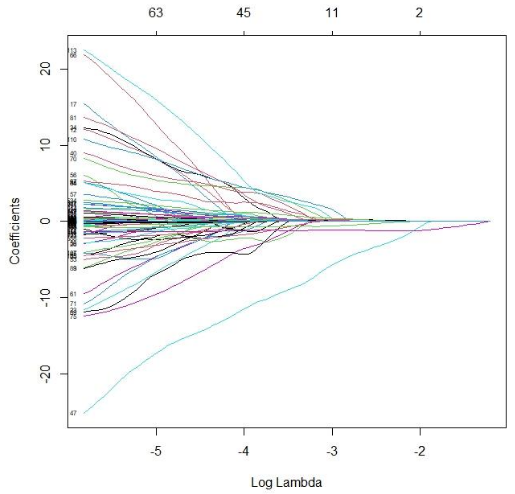


Lasso Regression for ADMCI Dataset

Boruta

Boruta feature selection algorithm is also used to select impactful predictors. Boruta reduces the number of predictors almost the same as Lasso for each dataset. The number of predictors after Boruta feature selection are:

- ADCTL Dataset: only **26** out of **429** predictors are available for training models
- ADMCI Dataset: only **7** out of **63** predictors are available for training models
- CTLMCI Dataset: only **16** out of **593** predictors are available for training models



Lasso Regression for MCICTL Dataset

Principle Component Analysis (PCA) & Boruta

A combined (or hybrid) feature selection method is also implemented in order to select the most impactful features from the training dataset which led to getting good prediction results and reduces overfitting. The combined PCA and Boruta feature selection algorithm reduced the number of predictors significantly. The number of predictors after Boruta feature selection are:

- ADCTL Dataset: only **8** out of **429** predictors are available for training models
- ADMCI Dataset: only **3** out of **63** predictors are available for training models
- CTLMCI Dataset: only **6** out of **593** predictors are available for training models

Training Models

Train Model Controller

10-fold cross-validation is used for training the datasets. In addition, datasets are centered and scaled for the training in order to reduce the biasing tendency towards predictors with higher values.

Classification Models

Multiple classification models are used in order to train each dataset and observed which model performs better. The testing dataset is fitted in the best classification model to get the optimum prediction. The used classification models are as follows:

1. Logistic Regression (LR)
2. Linear Discriminative Analysis (LDA)
3. Quadratic Discriminative Analysis (QDA)
4. Supper Vector Machine (SVM) [Linear/Gaussian Kernel]
5. K-Means Neighbor (KNN)
6. Naïve Bayes (NB)

Task 1: ADCTL Dataset Classification

The ADCTL dataset consists of **429 predictors** and **164 sample**. The task is to classify patient into two classes (**AD – Alzheimer Disease** and **CLT – Control**).
To perform the task, we must execute following steps:

- The ADCTL dataset is imported and after that the dimensionality of the dataset is calculated. It is noticed that there are **429 predictors** and **164 samples** which indicates the very high dimensionality and the possibility of overfitting.
- The balance info of the ADCTL is checked and it is found that there are **81 AD** and **83 CLT** classes. Therefore, it can be said that the dataset is balanced.
- **Variance Inflation Factor (VIF)** is used to find out predictors having co-linearity problem. In the case of ADCTL dataset, **110 predictors** have the co-linearity problems which are eliminated from the training dataset. As a result, the total number of predictors is **reduced to 320** after co-linearity checking.
- Predictors having **more than 75 %** correlation factor are removed from the training dataset. In the case of ADCTL dataset, **184 predictors** have **more than 75 %** correlation factor. Therefore, the total number of predictors is **decreased to 136.**

Task 1: ADCLT Dataset Classification

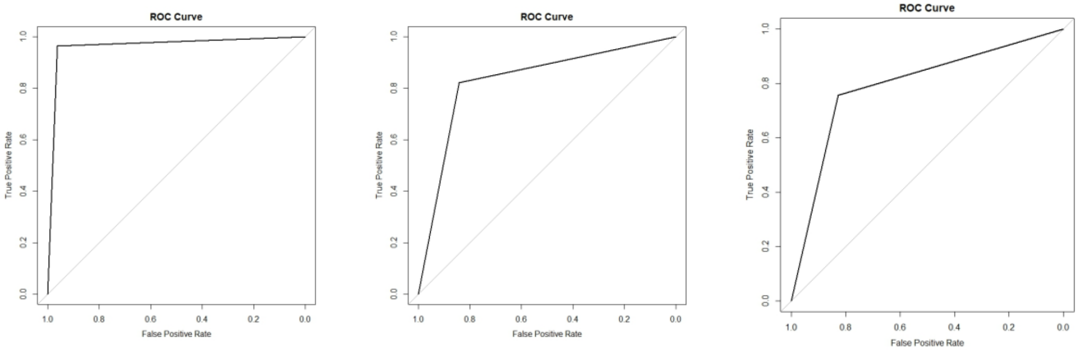
- Lasso Regression is used to select the most significant predictors which help to reduce overfitting. It makes the total number of predicators less (**only 24 predictors** are available for training models).
- Boruta feature selection method is also used beside Lasso and observed which one is performed well. Boruta feature selection algorithm provides **only 26 predictors** for training models which is higher than Lasso.
- In addition to Lasso and Boruta, a hybrid feature selection method is also used to get the most impactful features for training the models. The combination of PCA and Boruta provides **only 8 predictors** to train the models.
- Now, all the selected features from Lasso, Boruta, and the combination of PCA and Boruta are fitted into different classification models (e.g., logistic regression, support vector machine) separately and calculated the **AUC** and **MCC** score for each model.
- Finally, the best classification model is chosen according to the **highest AUC** and **MCC score**.

In the table-1, the performance on the ADCTL training dataset with different classification models are shown:

- The highest **AUC (0.963)** and **MCC (0.927)** scores are obtained with the combination of **Boruta** feature selection algorithm and **Quadratic Discriminative Analysis (QDA)** classification model.
- **Boruta** feature selection algorithm provides the best result among the three feature selection methods.

Table-1: Performance on the ADCTL Training Dataset

Models	Lasso		Boruta		PCA & Boruta	
	AUC	MCC	AUC	MCC	AUC	MCC
Logistic Regression	0.908	0.817	0.957	0.914	0.841	0.613
Linear Discriminative Analysis (LDA)	0.890	0.781	0.884	0.771	0.847	0.696
Quadratic Discriminative Analysis (QDA)	**0.951**	**0.902**	**0.963**	**0.927**	0.817	0.636
Supper Vector Machine (SVM - Linear)	0.902	0.805	0.945	0.890	**0.860**	**0.722**
Supper Vector Machine (SVM - Gaussian)	0.891	0.792	0.909	0.820	0.849	0.715
K-Means Neighbor (KNN)	0.830	0.666	0.866	0.737	0.836	0.679
Naïve Bayes (NB)	0.847	0.696	0.866	0.737	0.823	0.649



AUC Curve for the Best Prediction Models (ADCTL, ADMCI, MCICTL)

Task 2&3: ADMCI and MCICTL Dataset Classification

Table-2: Performance on the ADMCI Training Dataset

Models	Lasso		Boruta		PCA & Boruta	
	AUC	MCC	AUC	MCC	AUC	MCC
Logistic Regression	0.737	0.475	0.689	0.381	0.683	0.369
Linear Discriminative Analysis (LDA)	0.737	0.475	0.683	0.349	0.677	0.357
Quadratic Discriminative Analysis (QDA)	**0.831**	**0.663**	0.761	0.522	0.659	0.322
Supper Vector Machine (SVM - Linear)	0.743	0.487	0.720	0.441	0.689	0.381
Supper Vector Machine (SVM - Gaussian)	0.825	0.650	**0.827**	**0.656**	0.724	0.451
K-Means Neighbor (KNN)	0.767	0.534	0.777	0.557	**0.755**	**0.510**
Naïve Bayes (NB)	0.749	0.499	0.709	0.419	0.700	0.405

Table-3: Performance on the MCICTL Training Dataset

Models	Lasso		Boruta		PCA & Boruta	
	AUC	MCC	AUC	MCC	AUC	MCC
Logistic Regression	0.670	0.346	0.784	0.568	0.818	0.638
Linear Discriminative Analysis (LDA)	0.664	0.334	0.783	0.568	0.818	0.638
Quadratic Discriminative Analysis (QDA)	0.700	0.404	**0.876**	**0.755**	0.835	0.674
Supper Vector Machine (SVM - Linear)	0.685	0.370	0.792	0.584	0.818	0.638
Supper Vector Machine (SVM - Gaussian)	0.728	0.466	0.869	0.747	**0.852**	**0.709**
K-Means Neighbor (KNN)	**0.814**	**0.628**	0.854	0.708	0.829	0.662
Naïve Bayes (NB)	0.685	0.370	0.799	0.605	0.816	0.640

Table-4: Performance Summary on Training Datasets of the Best Model

Dataset	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Balanced Accuracy	AUC	MCC
ADCTL	0.9634	0.9638	0.9629	0.9638	0.9638	0.9634	**0.963**	**0.927**
ADMCI	0.8313	0.8222	0.8414	0.8505	0.8361	0.8318	**0.831**	**0.663**
MCICTL	0.8779	0.9000	0.8536	0.8709	0.8852	0.8768	**0.876**	**0.755**

In the table-2 & 3, the performance on the ADMCI and MCICTL training dataset with different classification models are shown:

- In the case of ADMCI, the highest **AUC (0.831)** and **MCC (0.663)** scores are obtained with the combination of **Lasso** feature selection algorithm and **Quadratic Discriminative Analysis (QDA)** classification model.

- In the case of MCICTL, the highest **AUC (0.876)** and **MCC (0.755)** scores are obtained with the combination of **Boruta** feature selection algorithm and **Quadratic Discriminative Analysis (QDA)** classification model.

- The **Lasso** feature selection algorithm provides the best result for ADMCI dataset whereas the **Boruta** feature selection algorithm provides the best result for the MCICTL dataset among the three feature selection methods.

- Table-4 shows the performance summary of each training dataset with their best classification model.