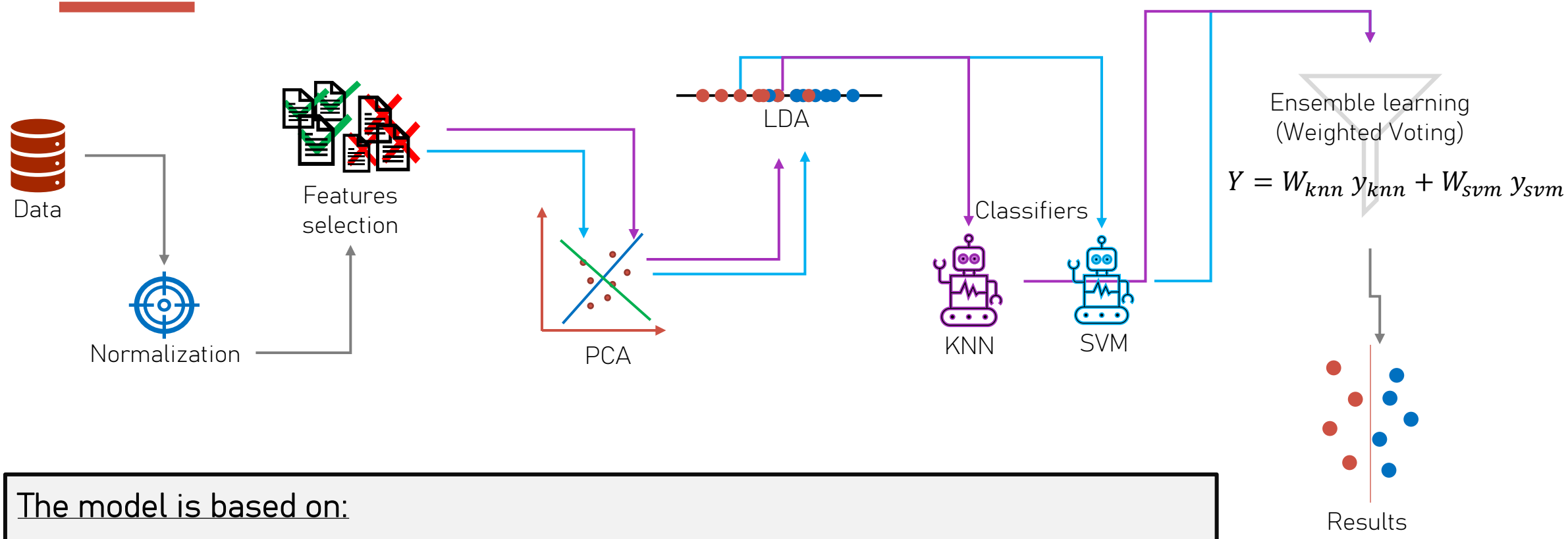
Classification challenge on Alzheimer's Disease using MRIs and Gene Expression data

STATISTICAL LEARNING AND DATA MINING

AGUSTIN CARTAYA

BASES OF THE MODEL



The model is based on:

1) Data normalization
2) Features selection
3) PCA
4)LDA
5) KNN - SVM classifiers
6) Ensemble learning (Weighted Voting)

Feature selection

The feature selection was performed using the selectKBest method from the sklearn library.

For each dataset, K features were selected using the combination of two scoring functions:

- **f_classif [f]:** Calculates the ANOVA F-value and p-value for each feature, indicating their statistical significance in predicting the target variable.
- **mutual_info_classif [m]:** Relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances.

Then, the selected features from both functions were combined.

The number of features, K, depends on each classifier (SVM, KNN) and was obtained by testing different combinations for each dataset and classifier.

For the:

- ADCTL dataset k = 358 = (f = 1 U m = 358) features were selected for the knn and k = 247 = (f = 1 U m = 247) for the svm

- ADMCI dataset k = 43 = (f = 6 U m = 43) features were selected for the knn and k = 47 = (f = 53 U m = 50) for the svm

- MCICTL dataset k = 172 = (f = 1 U m = 172) features were selected for the knn and k = 172 = (f = 1 U m = 172) for the svm

PCA / LDA

After feature selection, the top n principal components were extracted. The PCA function from the sklearn library was used to calculate the principal components.

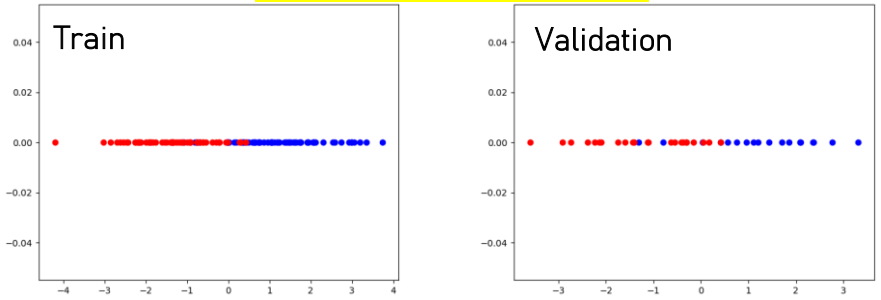
The number of principal components, n, was determined simultaneously with k, aiming to maximize the accuracy on the validation dataset. This number depends on the specific dataset and classifier used.

For the:

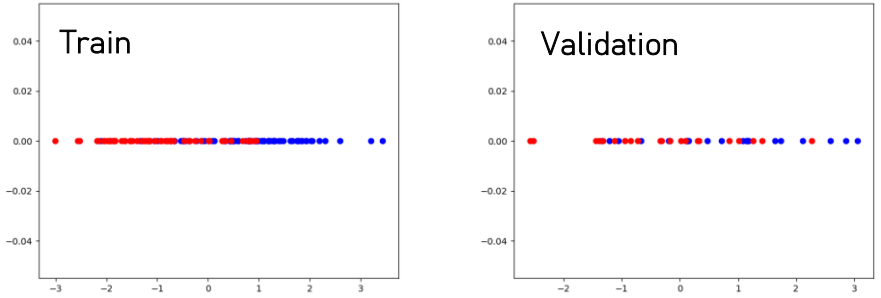
- ADCTL dataset: n = 9 for the knn classifier and n = 119 for the svm classifier.
- ADMCI dataset: n = 19 for the knn classifier and n = 50 for the svm classifier.
- MCICTL dataset: n = 61 for the knn classifier and n = 51 for the svm classifier.

After obtaining the principal components, Linear Discriminant Analysis was applied to them:

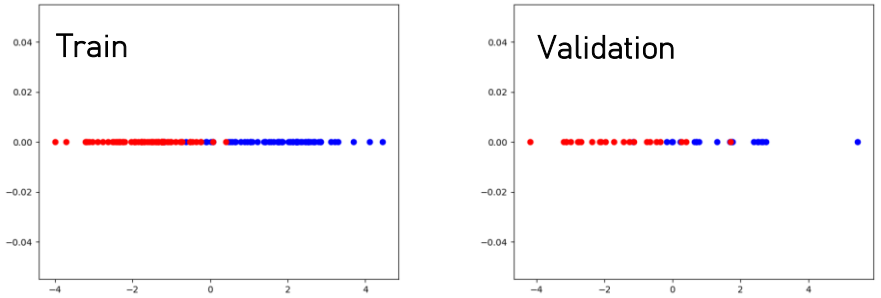
LDA (ADCTL dataset)



LDA (ADMCI dataset)



LDA (MCICTL dataset)



Results are obtained with the feature selection and PCA corresponding to the knn

Ensemble learning

After applying LDA, the classifiers were trained, following an Ensemble Learning model with Weighted Voting for classification.

The parameters of the classifiers were chosen using grid search. These parameters depend on the dataset.

For the:

- ADCTL dataset: n_neighbors = 3 and p = 1 for the knn classifier, and C = 0.1, degree = 1, and kernel = poly for the svm classifier.
- ADMCI dataset: n_neighbors = 2 and p = 1 for the knn classifier, and C = 1.0, degree = 1, and kernel = poly for the svm classifier.
- MCICTL dataset: n_neighbors = 3 and p = 1 for the knn classifier, and C = 0.1, degree = 1, and kernel = poly for the svm classifier.

After obtaining the classification percentages from the classifiers, a final percentage was calculated by applying a linear combination of the results, with weights assigned to each classifier.

The weights of each classifier were calculated based on the AUC (Area Under the Curve) of the validation test (val) for each dataset. For example, for the ADCTL dataset, the weights were calculated as follows:

$$w_{knn} = \frac{knn_{val_{auc}}}{knn_{val_{auc}} + svm_{val_{auc}}} = 0.495$$

$$w_{svm} = \frac{svm_{val_{auc}}}{knn_{val_{auc}} + svm_{val_{auc}}} = 0.505$$

Thus, the linear combination of the classifiers is as follows:

$$Y = 0.495\, y_{knn} + 0.505\, y_{svm}$$

Results

Performance on the 75% of the training datasets (data used to train the models)

DATASET	Acc	Sens	Spec	Prec	F1	AUC	MCC	BA
ADCTL	1	1	1	1	1	1	1	1
ADMCI	0.914728682	0.918032787	0.911764706	0.903225806	0.910569106	0.982160077	0.829198	0.914898746
MCICTL	0.96124031	0.970149254	0.951612903	0.955882353	0.962962963	0.996870486	0.922428571	0.960881078

Performance on the 25% of the training datasets (data used as validation, and not used to train the models)

DATASET	Acc	Sens	Spec	Prec	F1	AUC	MCC	BA
ADCTL	0.902439024	0.869565217	0.944444444	0.952380952	0.909090909	0.990338164	0.808174392	0.907004831
ADMCI	0.697674419	0.666666667	0.727272727	0.7	0.682926829	0.792207792	0.394794855	0.696969697
MCICTL	0.930232558	0.913043478	0.95	0.954545455	0.933333333	0.965217391	0.861173393	0.931521739

Performance on the 100% of the training datasets

DATASET	Acc	Sens	Spec	Prec	F1	AUC	MCC	BA
ADCTL	1	1	1	1	1	1	1	1
ADMCI	0.88372093	0.865853659	0.9	0.8875	0.87654321	0.968157182	0.766893515	0.882926829
MCICTL	0.918604651	0.922222222	0.914634146	0.922222222	0.922222222	0.982655827	0.836856369	0.918428184

Performance on 5-fold cross validation in the training set

DATASET	Acc	Sens	Spec	Prec	F1	AUC	MCC	BA
ADCTL	0.847537879	0.865441176	0.822745098	0.854879623	0.85539549	0.924329287	0.700672392	0.844093137
ADMCI	0.686554622	0.659649123	0.72128655	0.689615385	0.66755937	0.719653509	0.38335512	0.690467836
MCICTL	0.790252101	0.814524114	0.763636364	0.780657939	0.793093573	0.874776575	0.578214929	0.789080239