REPORT PART 1

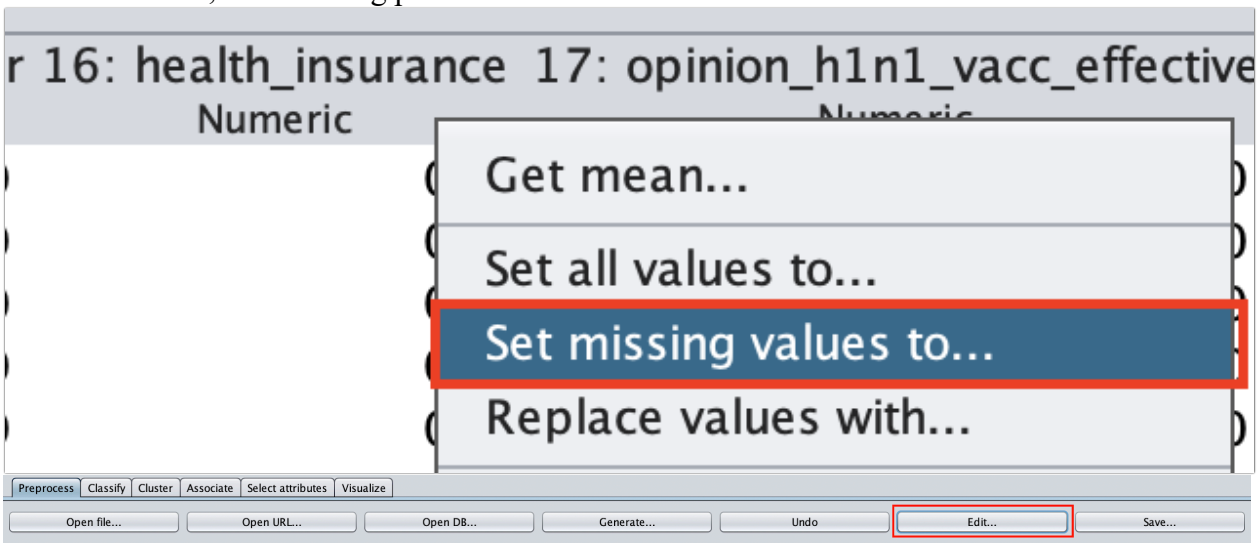
Preprocessing:

In order to preprocess the dataset, different preprocessing methods were applied to analyse the probabilities of H1H1 vaccine and seasonal vaccine.

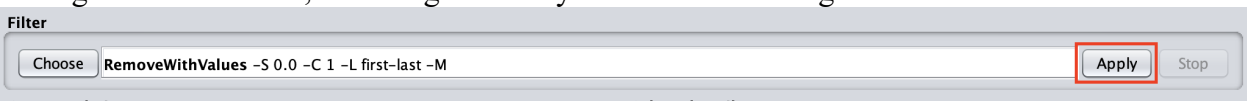
The following attributes were deleted because they are related to vaccines that are not being analysed or have more than half of the rows with null values:

- Attributes deleted for both analysis: respondent_id, employment_industry, employment_occupation.
- Attributes deleted for the H1N1 analysis: doctor_recc_seasonal, opinion_seas_vacc_effective, opinion_seas_risk, opinion_seas_sick_from_vacc.
- Attributes deleted for the seasonal vaccine analysis: doctor_recc_h1n1, opinion_h1n1_vacc_effective, opinion_h1n1_risk, opinion_h1n1_sick_from_vacc, h1n1_concern', h1n1_knowledge, health_insurance, census_msa.

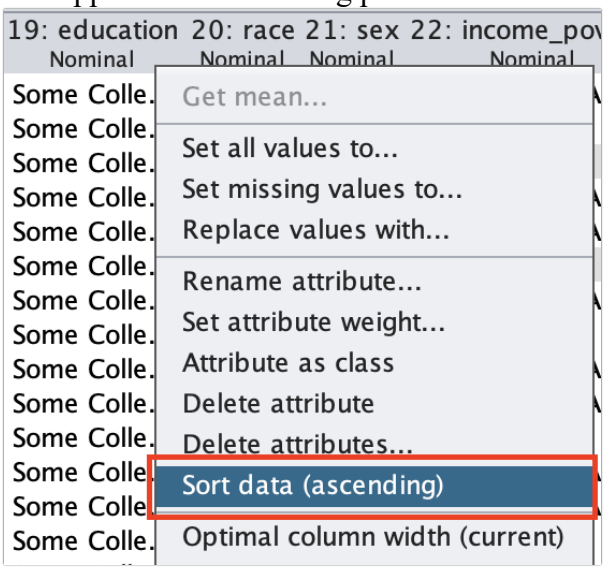
The health insurance attribute is important for H1N1 analysis. Because of that, missing values were set with -1, not affecting predictions that have health insurance information.



The function RemoveWithValues with the parameter matchMissingValues set True was applied to all numeric fields, deleting instances containing null values. This function was applied due to the high amount of data, not being necessary to calculate missing values for each attribute.



Due to limitations on Weka, the RemoveRange function was used in order to remove null values in nominal attributes. The dataset was ordered by each nominal attribute and then the function was applied. The following process was done for each nominal attribute:



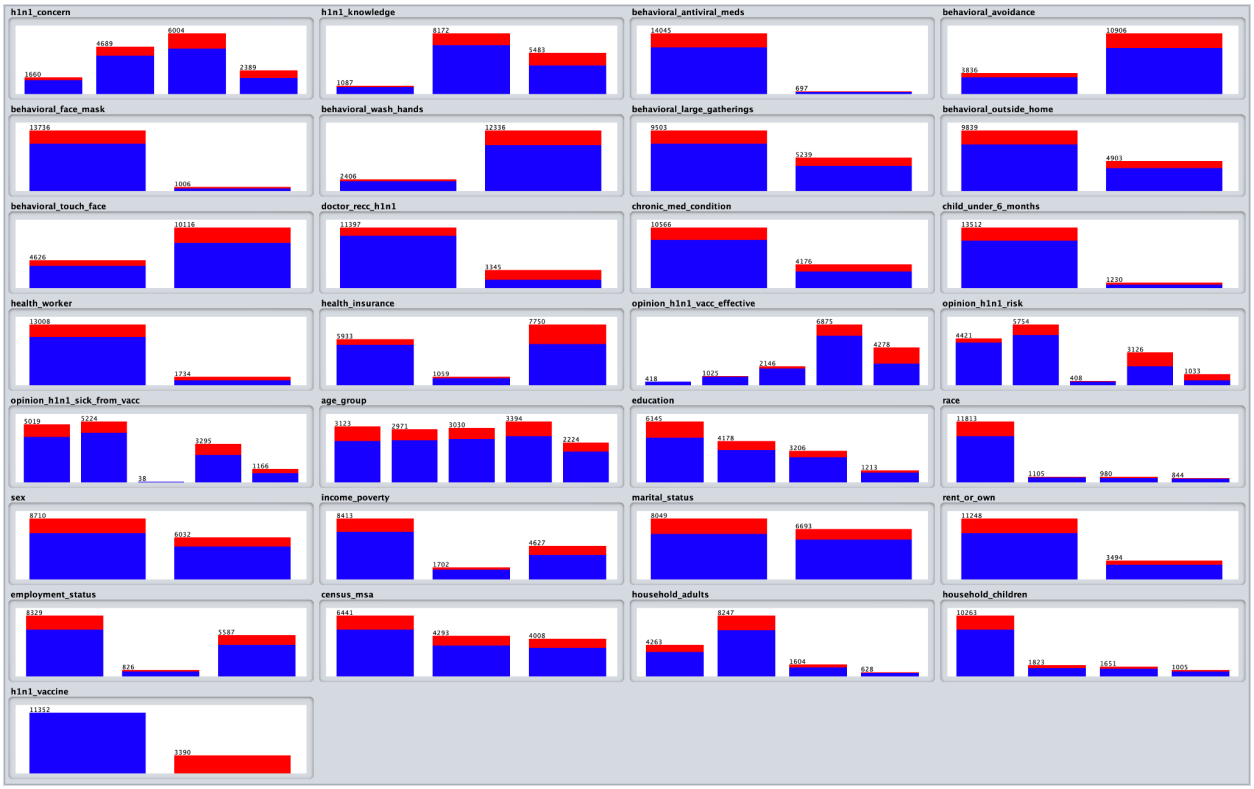
Parameter InstancesIndices was replaced by the formula (Total Instances - Missing values in attribute) - last.

Finally, all numeric attributes were converted to nominal using the function NumericToNominal:

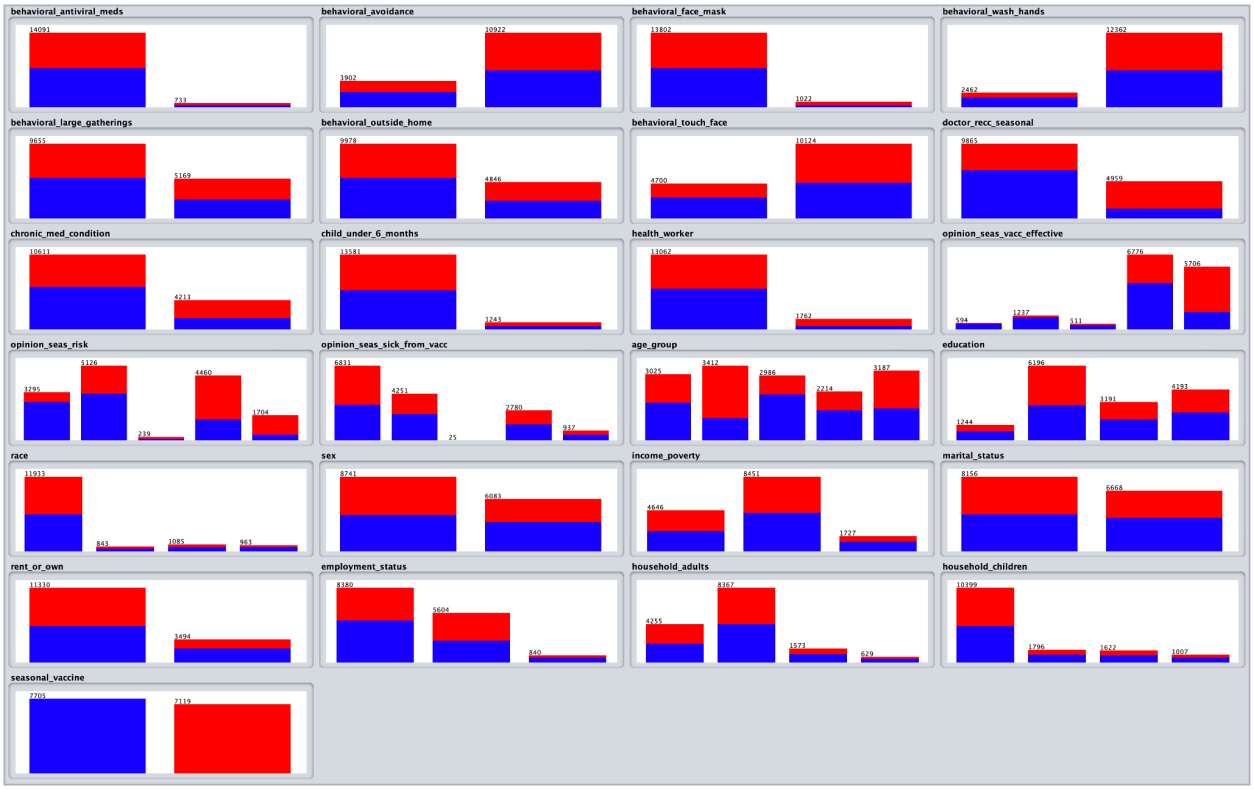
EDA

The dataset is composed of categorical attributes. Due to the nature of the dataset, bar charts are one of the best ways to perform descriptive analysis. EDA was done for the H1N1 and seasonal datasets, where the red color represents who took the shot and blue represents who didn't take the shot.

H1N1 vaccine dataset:



EDA Seasonal vaccine dataset:



Training, testing and validation sets

In order to split the dataset, the function RemovePercentage was used to create the training and validation dataset.

To create the training dataset, 25% of the data was removed, leaving 75% for training. Change the percentage parameter to 25.

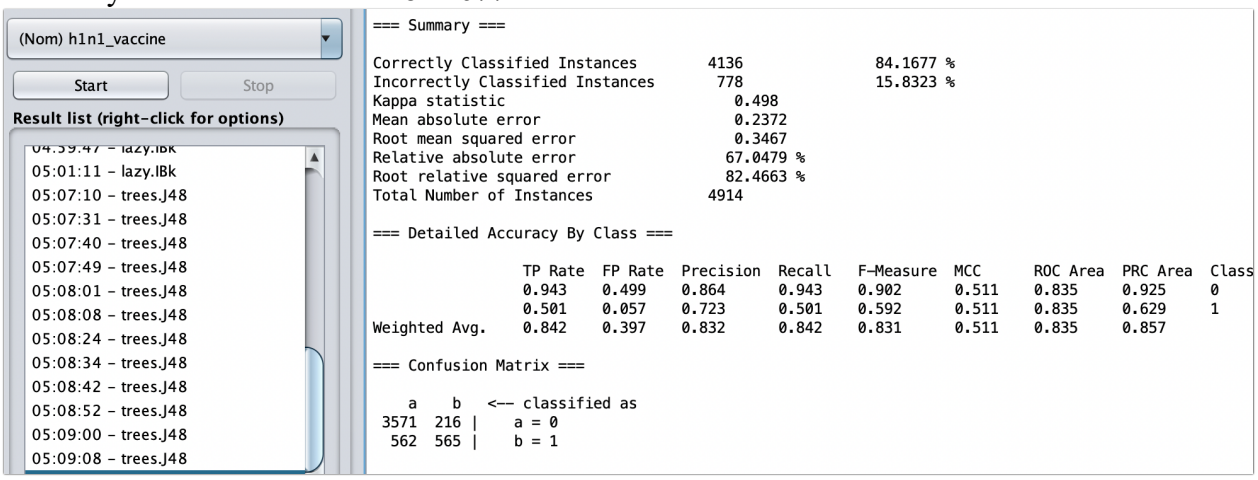
To create the validation dataset, the previous arff was loaded and 75% of the data was removed, leaving 25% for test. Change the percentage parameter to 25 and invertSelection to True.

The test dataset comes from a different file called test_set_features. This dataset is used by the driven data to analyse the results given.

Classifier(s) used / Association / Clustering

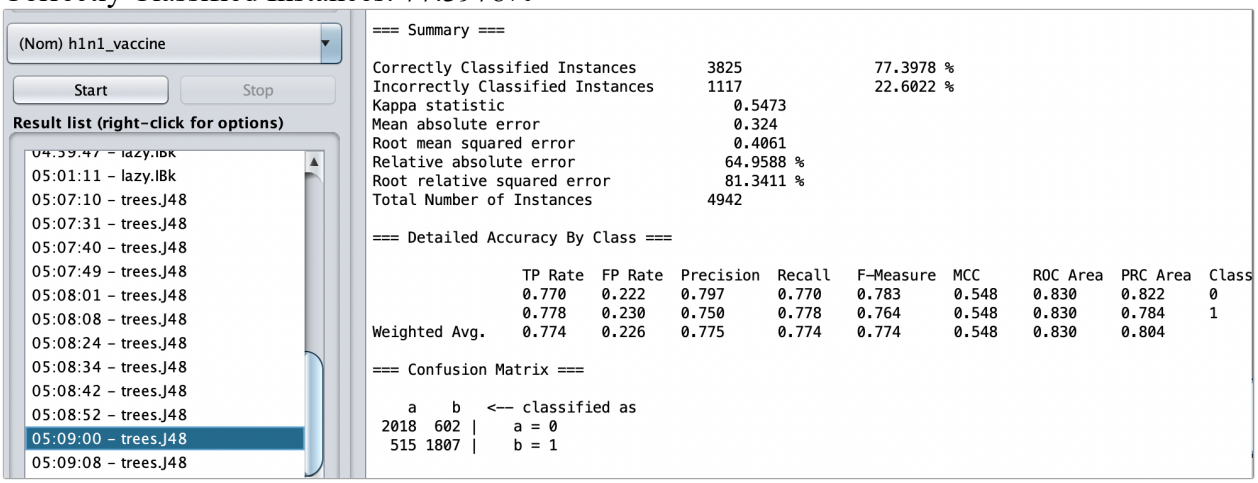
DecisionTree H1N1:

Number of leaves: 520
Size of the tree: 788
Correctly Classified Instances: 84.1677%



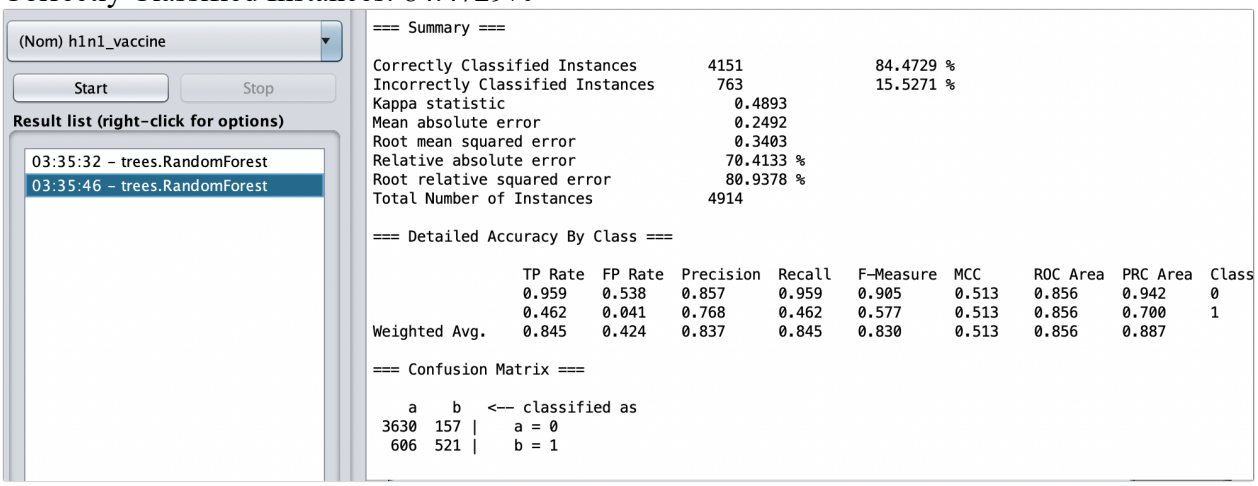
DecisionTree seasonal:

Number of leaves: 601
Size of the tree: 926
Correctly Classified Instances: 77.3978%



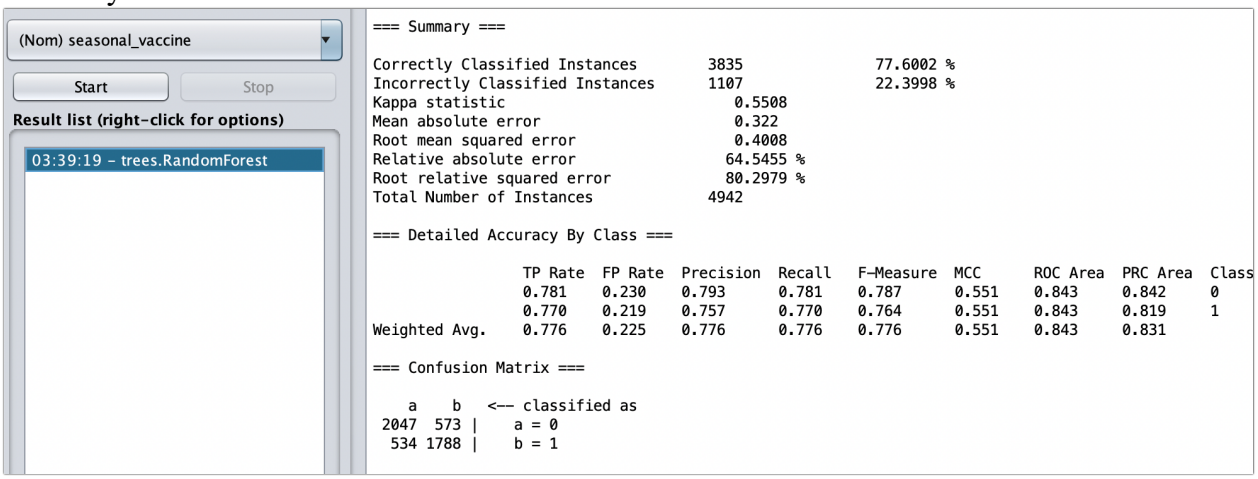
Random forest H1N1:

Parameters Used: Iterations = 500, depth = 70
Correctly Classified Instances: 84.4729%



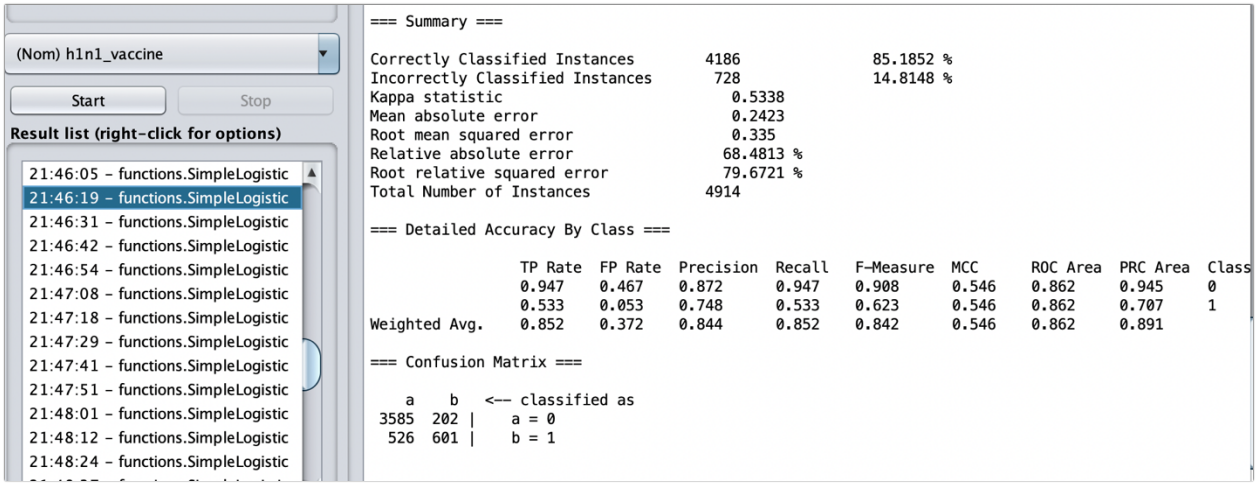
Random forest seasonal:

Parameters Used: Iterations = 500, depth = 70
Correctly Classified Instances: 77.6002%



LogisticRegression H1N1:

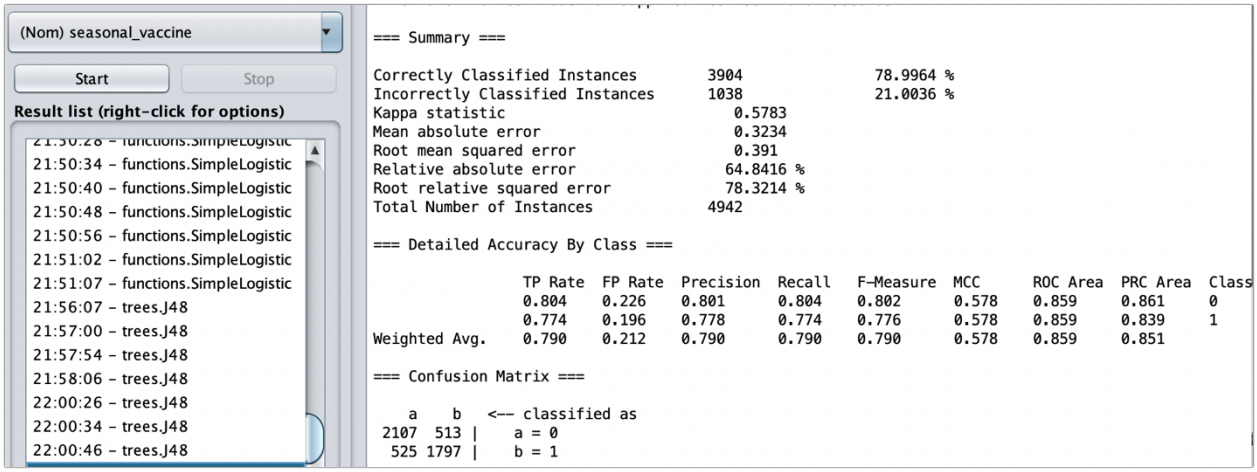
Parameters Used: useAIC = True, weightTrimBeta = 0.1



Correctly Classified Instances: 85.1852%

LogisticRegression seasonal:

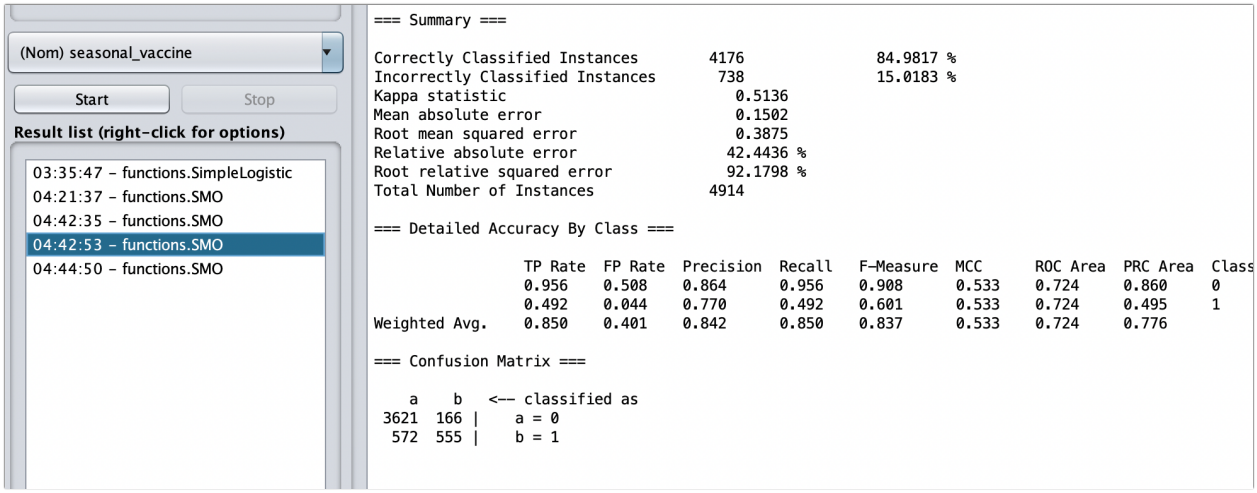
Parameters Used: useAIC = True, weightTrimBeta = 0.1



Correctly Classified Instances: 78.9964%

Support vector machine H1N1:

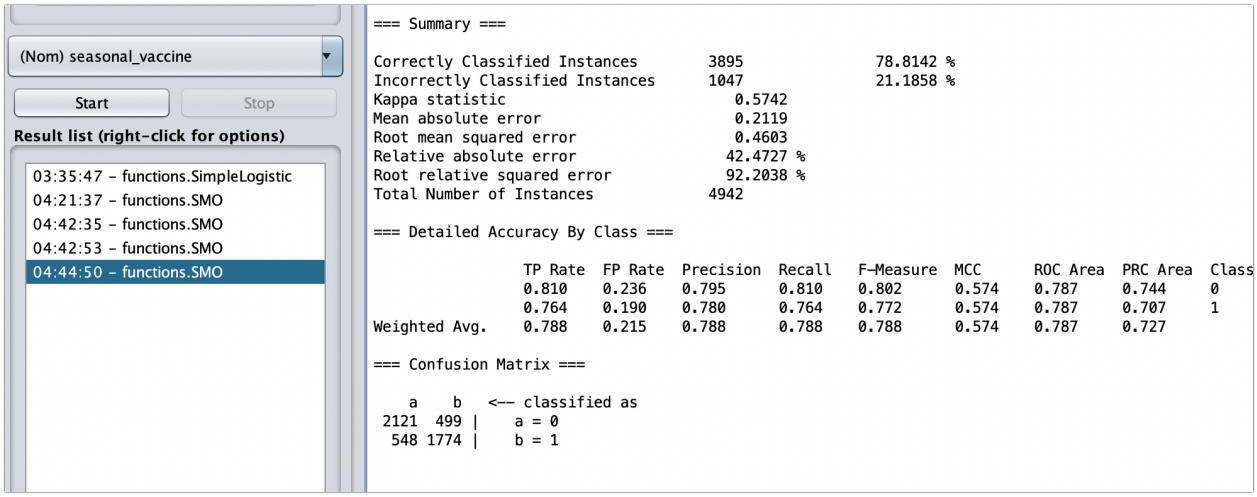
Parameters Used: Kernel = RBF



Correctly Classified Instances: 84.9817%

Support vector machine seasonal:

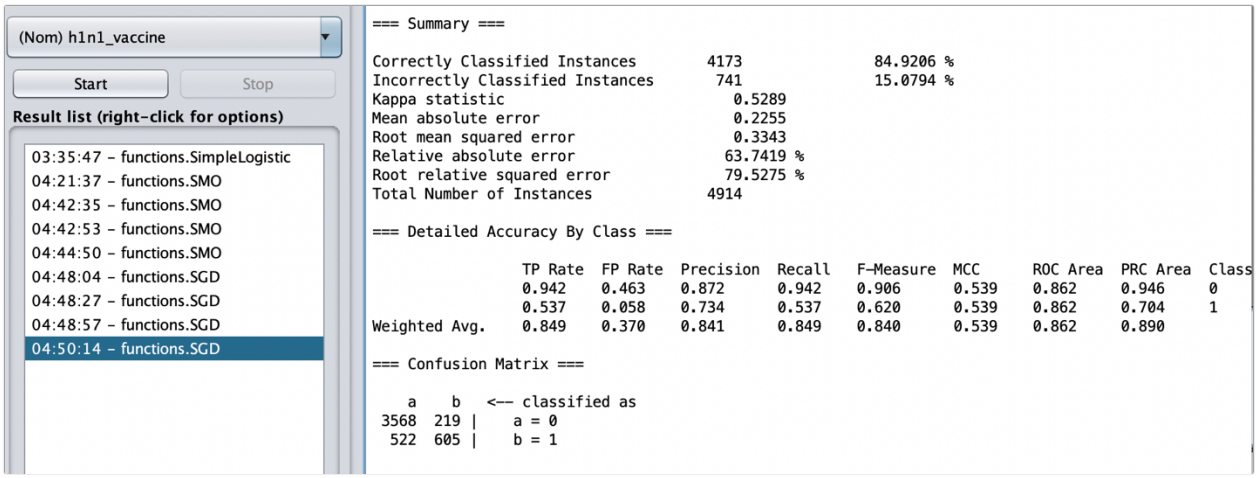
Parameters Used: Kernel = RBF



Correctly Classified Instances: 78.8142%

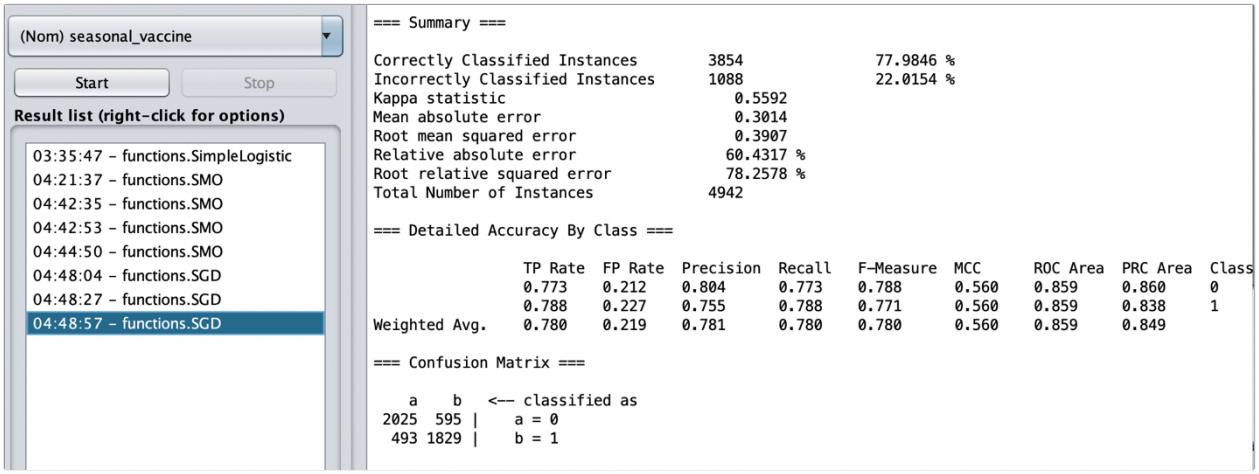
Stochastic gradient descent H1N1:

Parameters Used: Kernel = Epochs = 500, lossFunction=Log Loss
Correctly Classified Instances: 84.9206%



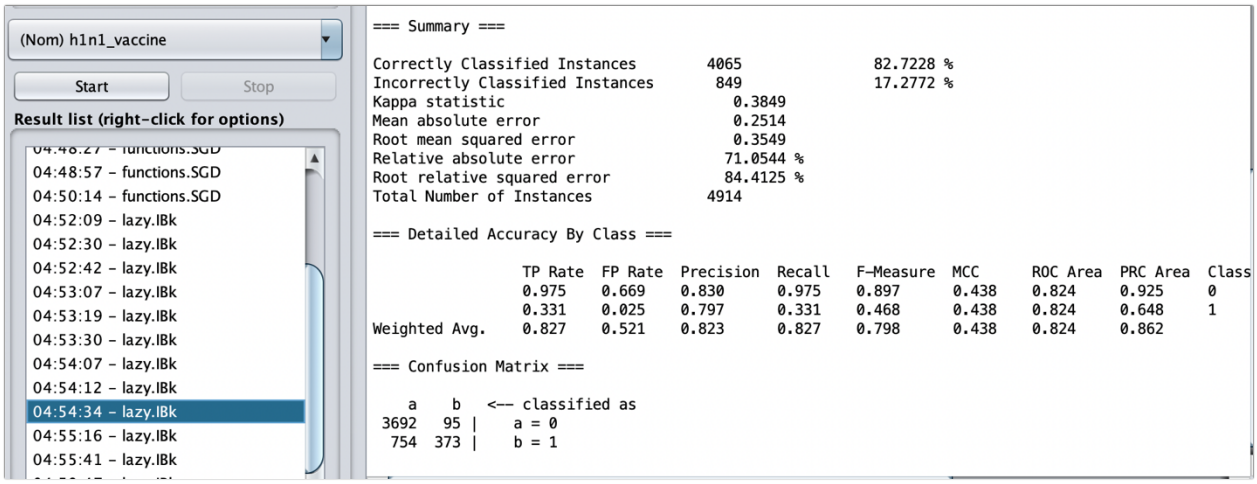
Stochastic gradient descent seasonal:

Parameters Used: Kernel = Epochs = 500, lossFunction=Log Loss
Correctly Classified Instances: 77.9846%



KNN classifier H1N1:

Parameters Used: Neighbours = 10, Algorithm = LinearNNSearch, DistanceFunction = Manhattan Distance
Correctly Classified Instances: 82.7228%



KNN classifier seasonal:

Parameters Used: Neighbours = 10, Algorithm = LinearNNSearch, DistanceFunction = Manhattan Distance
Correctly Classified Instances: 76.4063%