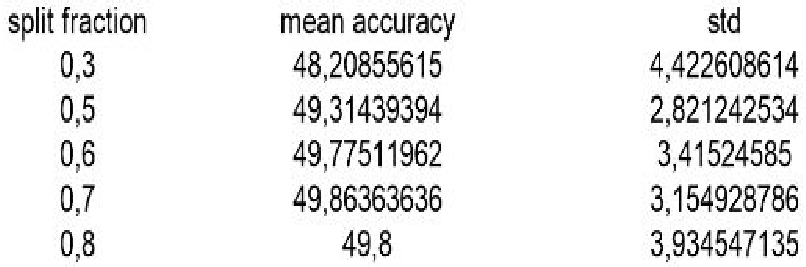
Naive bayes classification

The justification of the naive assumption

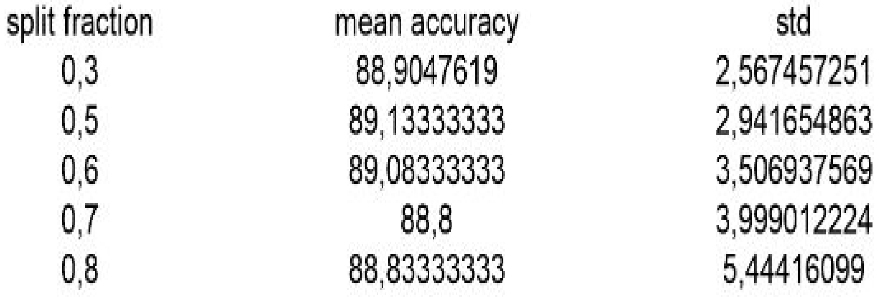
The naive bayes assumption assumes that every observed features are independent of each other given the target class. This assumption is justified if the underlying feature of the data are truly independent. For example there are some correlation between age | length | vocabulary capacity, age and vocabulary capacity is not independent if we analyse them together but there truly independent if the target class is fixed. Given a fixed target class 'length' then P(age|length) and P(voc|length) are independent.

As described before the NBC is a very good and robust classifier if the observed features are probabilistic independent of each other. In the cases when that independence is not hold then a high rate of misclassification can be observed. That can occur in text classification tasks when the probability to find one word is not independent of finding another word in the text. for example in New York, so is New and York dependent of each other in that context. In those cases the NBC may performs poorly as observed in the Classification of vowels dataset:



split fraction	mean accuracy	std
0,3	48,20855615	4,422608614
0,5	49,31439394	2,821242534
0,6	49,77511962	3,41524585
0,7	49,86363636	3,154928786
0,8	49,8	3,934547135

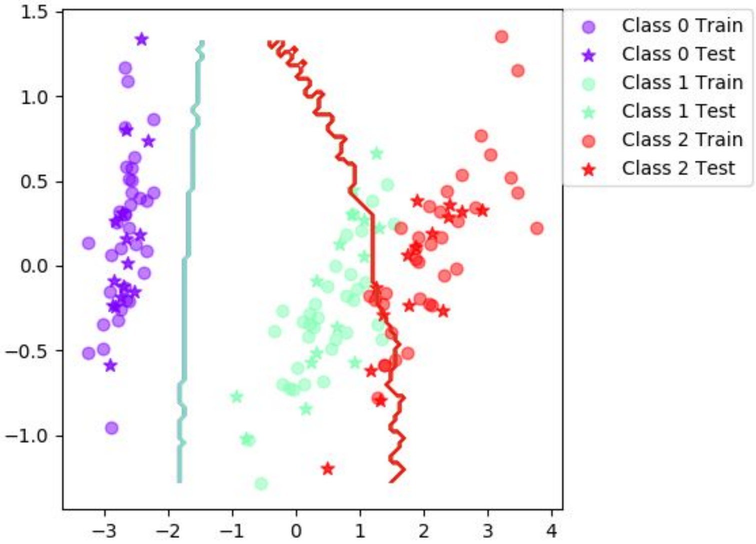
VOWEL DATA



split fraction	mean accuracy	std
0,3	88,9047619	2,567457251
0,5	89,13333333	2,941654863
0,6	89,08333333	3,506937569
0,7	88,8	3,999012224
0,8	88,83333333	5,44416099

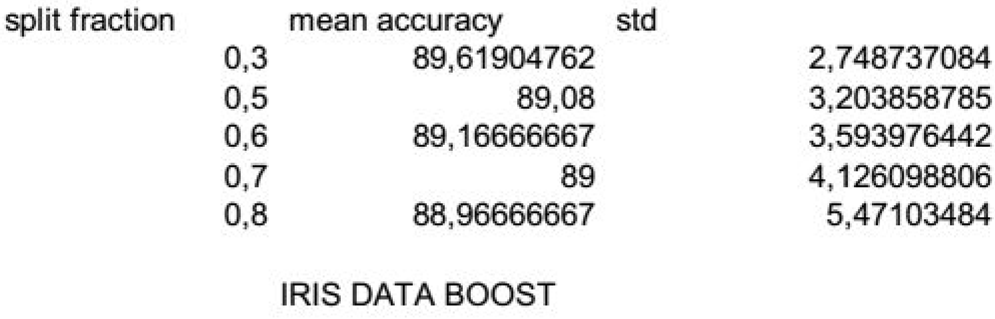
IRIS DATA

Improvement regarding the decision boundary



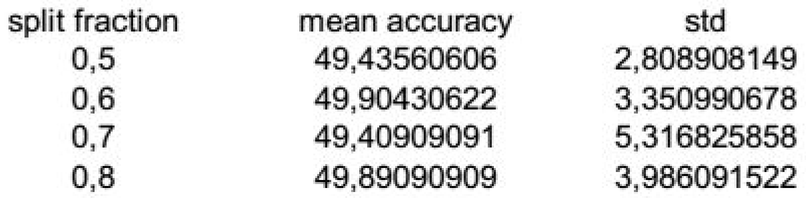
We can see that the decision boundary between the class 1 and class 2 is very difficult to determine. It seems like there are some occurence of overfitting and the presence of outliers. To improve the accuracy, we could change the classifier and use for example *the support vector machine **(SVM)** with non-linear kernel to transform the data and make it more separable* (eventually with slack variables). We can have ***more data*** to find out exactly if the outliers observed are only single errors or if they really represent a pattern in the dataset. We could also try to transform the data by *the **PCA** method* and find out if there are some dimensionality in which the data are more separable and have a higher predictive accuracy.

Boosting



split fraction	mean accuracy	std
0,3	89,61904762	2,748737084
0,5	89,08	3,203858785
0,6	89,16666667	3,593976442
0,7	89	4,126098806
0,8	88,96666667	5,47103484

IRIS DATA BOOST

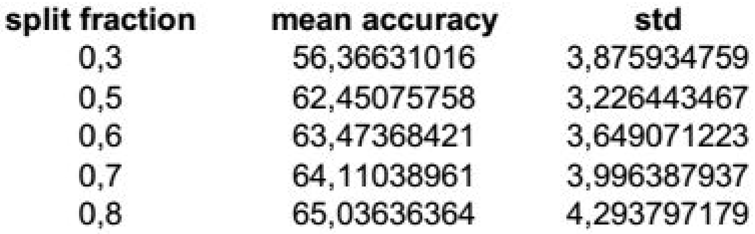


split fraction	mean accuracy	std
0,5	49,43560606	2,808908149
0,6	49,90430622	3,350990678
0,7	49,40909091	5,316825858
0,8	49,89090909	3,986091522

VOWEL DATA BOOST

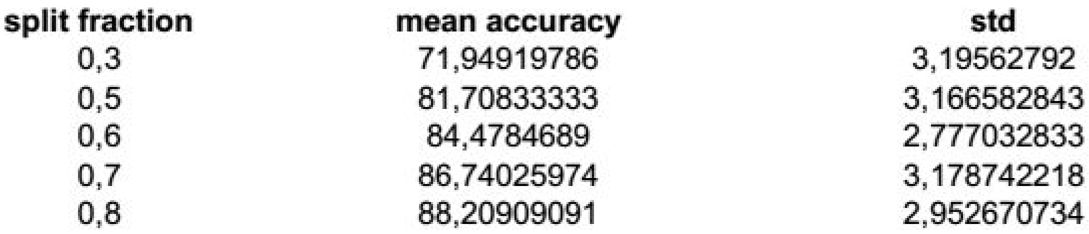
We can observe no improvement on the vowel dataset prediction accuracy due to boosting, only between 0 to 0.3 procentual improvement. That change is not big enough to be considered as an improvement. But on the Iris dataset we have 1 % procentual improvement that is slightly better than without boosting. The reason behind the non-improvement on the vowel dataset may be the bayes assumption that is not really fitted for the given data and that the learners is not weak enough the change due to weighting.

Decision Trees



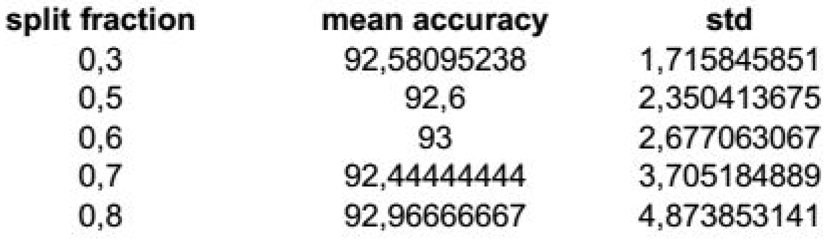
split fraction	mean accuracy	std
0,3	56,36631016	3,875934759
0,5	62,45075758	3,226443467
0,6	63,47368421	3,649071223
0,7	64,11038961	3,996387937
0,8	65,03636364	4,293797179

VOWEL Decision tree



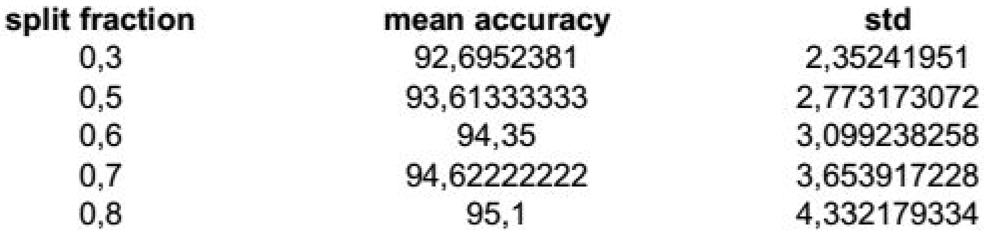
split fraction	mean accuracy	std
0,3	71,94919786	3,19562792
0,5	81,70833333	3,166582843
0,6	84,4784689	2,777032833
0,7	86,74025974	3,178742218
0,8	88,20909091	2,952670734

VOWEL DECISION TREE BOOST



split fraction	mean accuracy	std
0,3	92,58095238	1,715845851
0,5	92,6	2,350413675
0,6	93	2,677063067
0,7	92,44444444	3,705184889
0,8	92,96666667	4,873853141

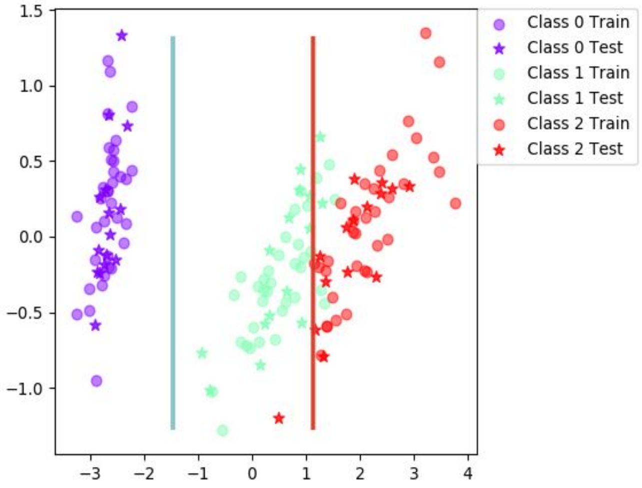
IRIS DECISION TREE



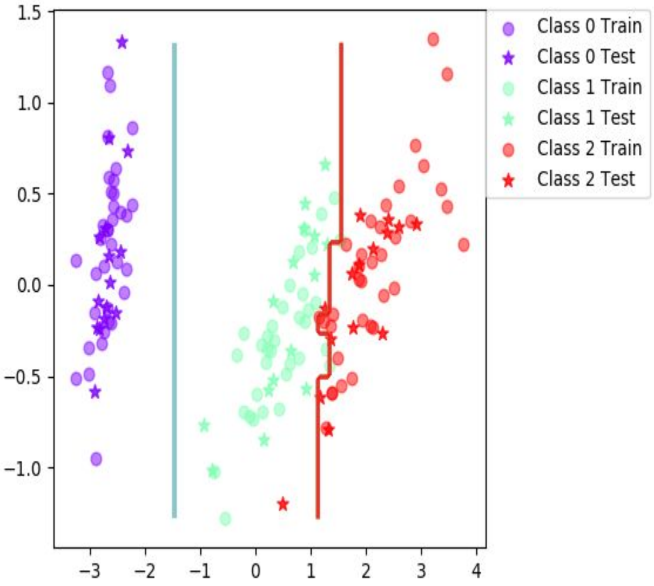
split fraction	mean accuracy	std
0,3	92,6952381	2,35241951
0,5	93,61333333	2,773173072
0,6	94,35	3,099238258
0,7	94,62222222	3,653917228
0,8	95,1	4,332179334

IRIS DECISION TREE BOOST

<u>IRIS DATA BOUNDARY</u>

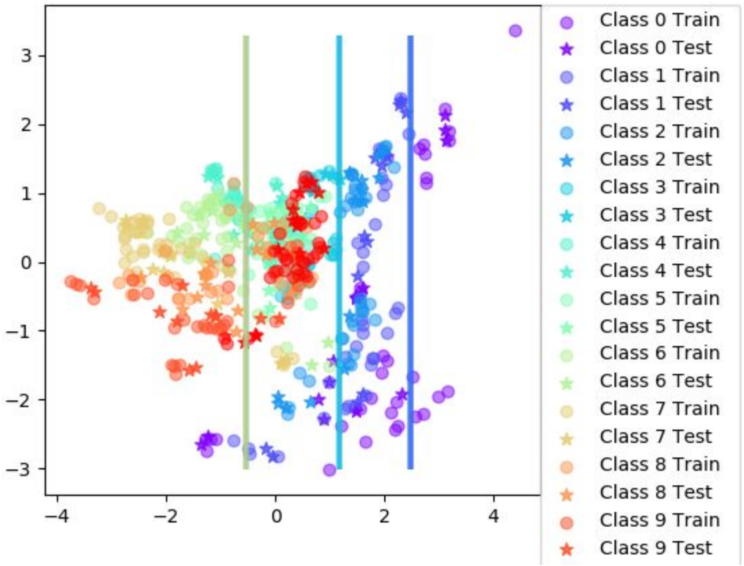


split 0.7 decision tree boundary without boosting

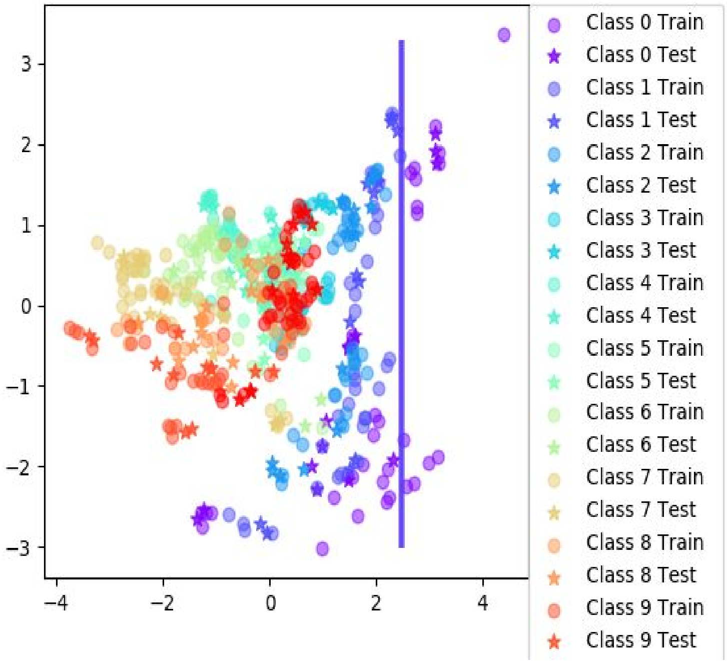


split 0.7 decision tree boundary wich boosting

VOWEL DATA BOUNDARY



split 0.7 decision tree boundary without boosting



split 0.7 decision tree boundary with boosting