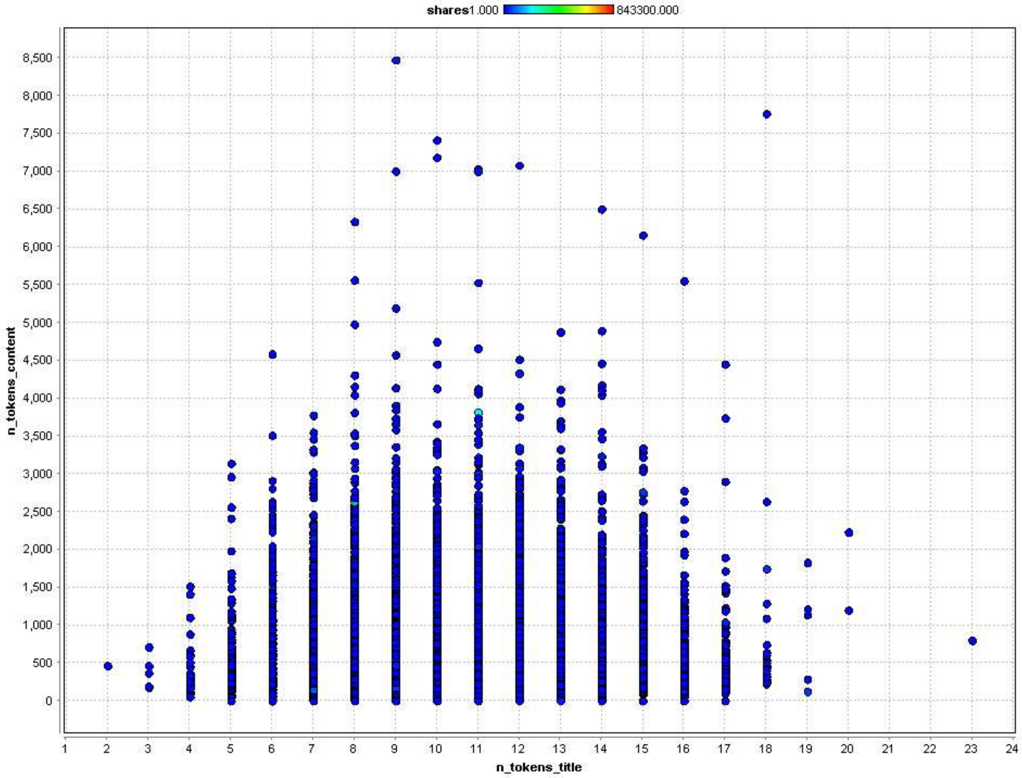
Practical task: Article popularity prediction on „Online Online News Popularity Data Set" in RapidMiner following CRISP guidelines

Business Understanding:

Authors and project managers can concentrate on producing content that consumers appreciate.

Data Understanding:

39797 instances / articles. 61 attributes.
Both integer, continuous valued and binominal attributes.



Data Preparation

Useless attributes removed: *timedelta*, *url*

Two attributes removed because of correlation with other attributes,
or just not giving any informaton:
n_non_stop_unique_tokens highly correlated with *n_unique_tokens*.
n_non_stop_words only zero.

Integer attributes converted into continuous valued attributes to better work with modelling tools.

Target variable *shares*: Number of online shares. Made into binominal using threshold: > 1400: Popular article, else not popular

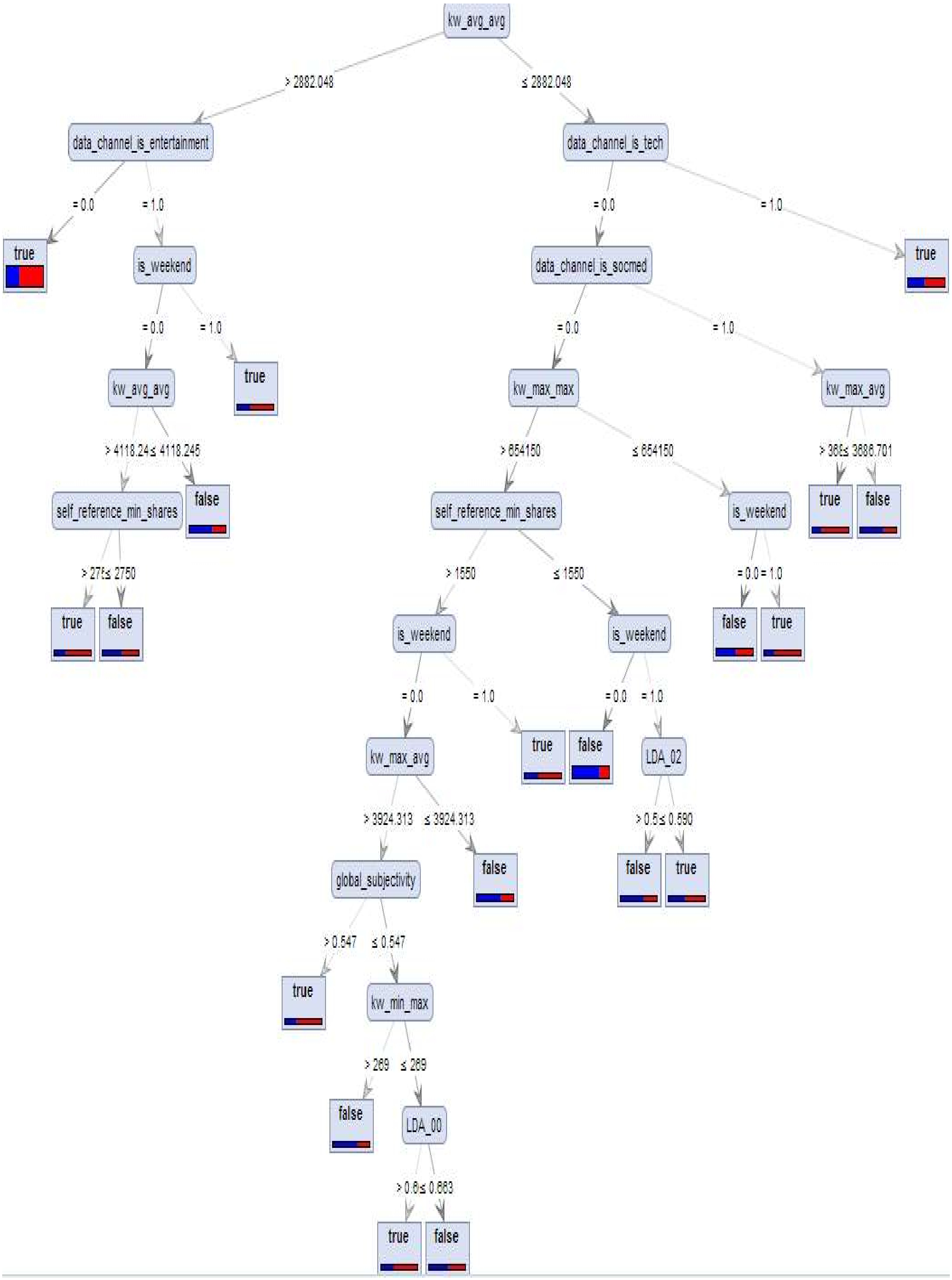
Modelling

Goal: Predict if article is popular using X-validation training / validation split.
Using different model building techniques.

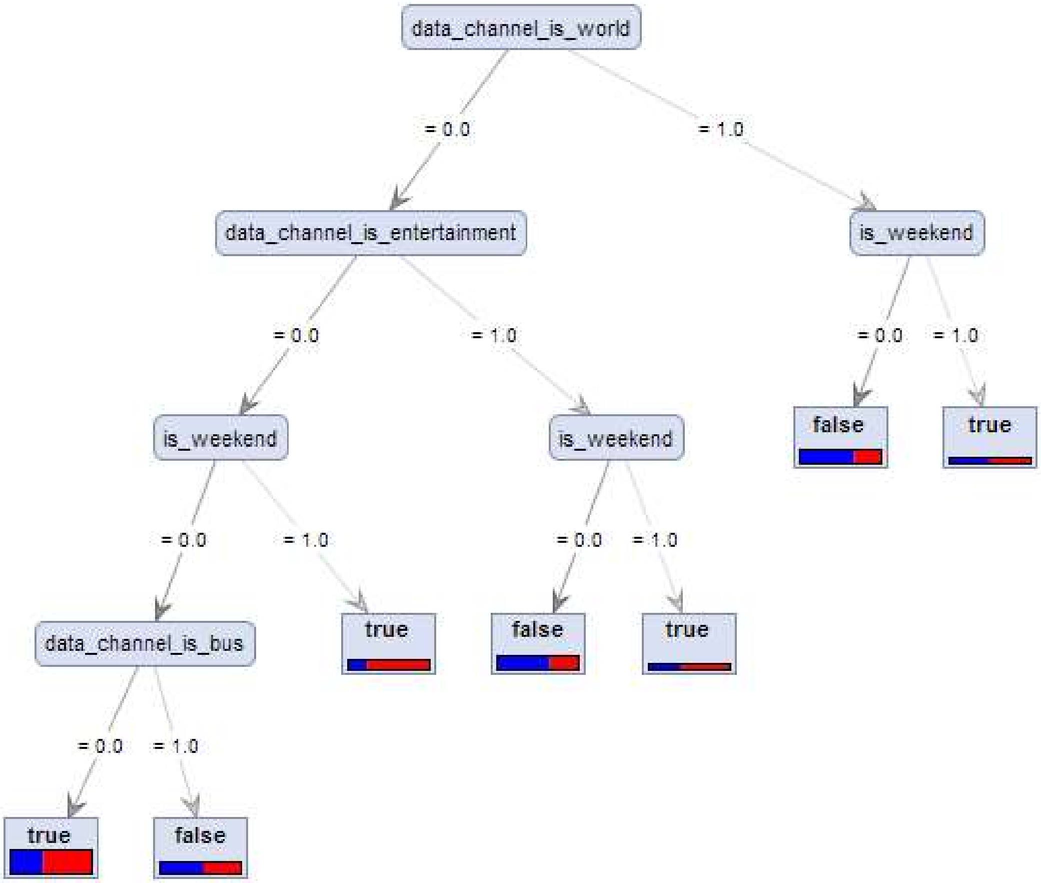
Decision Tree

Vizuaisation of decision tree on (all equal configurations)

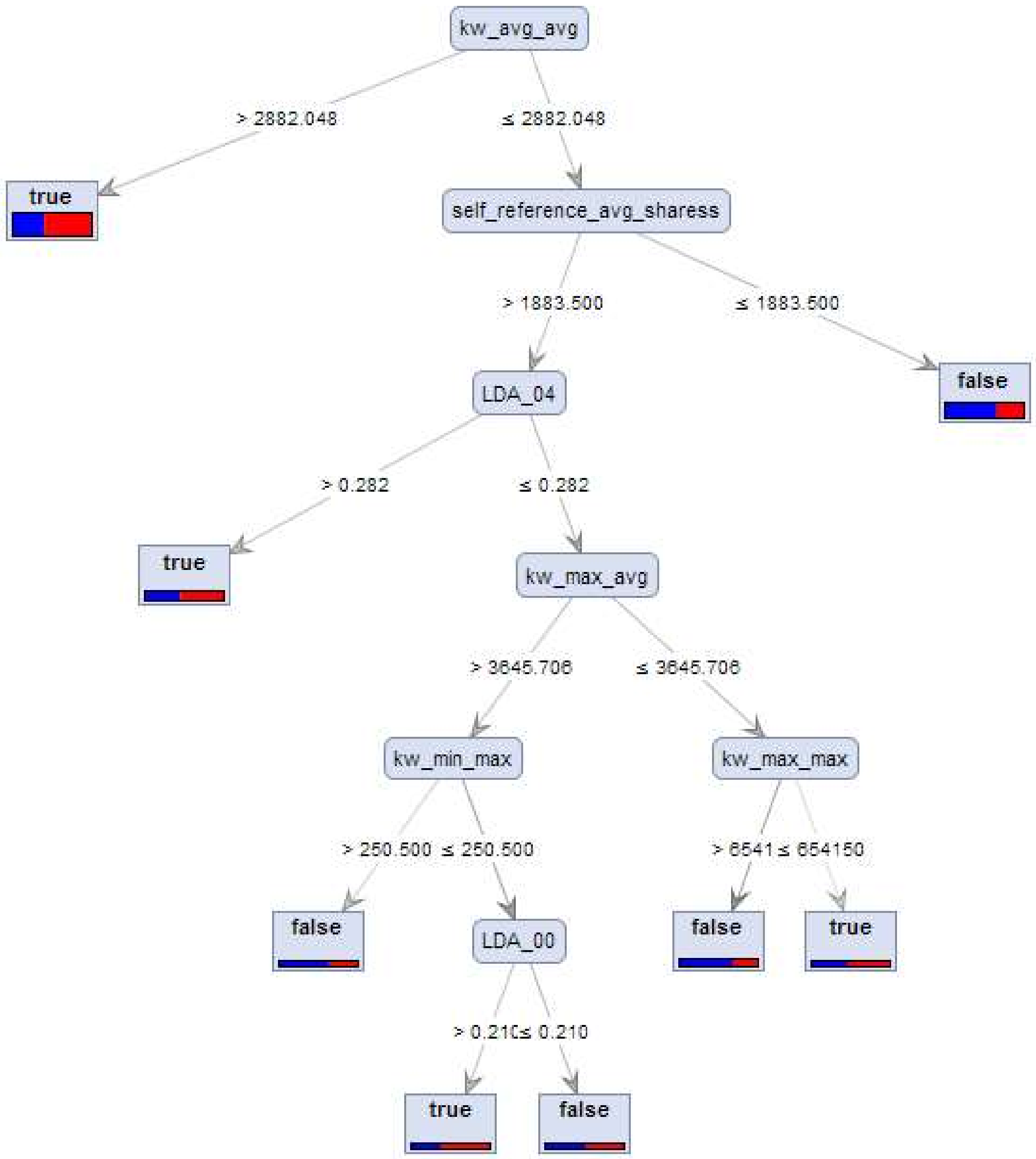
Full data



Only binominal attributes



Only continous attributes

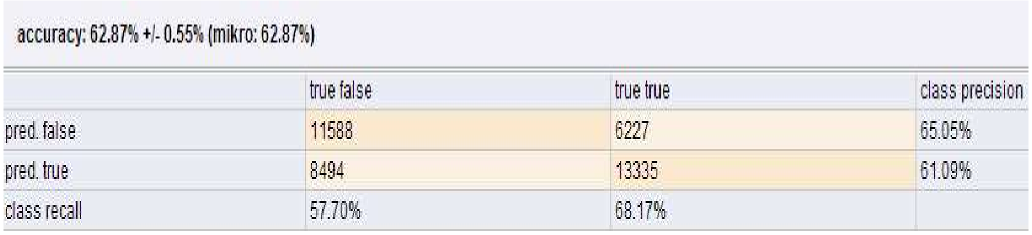


k-NN
Naïve Bayes

Evaluation and Depoyment

Classification accuracy for the different methods compared, using X-validation with apply model and classification performance. All the data.

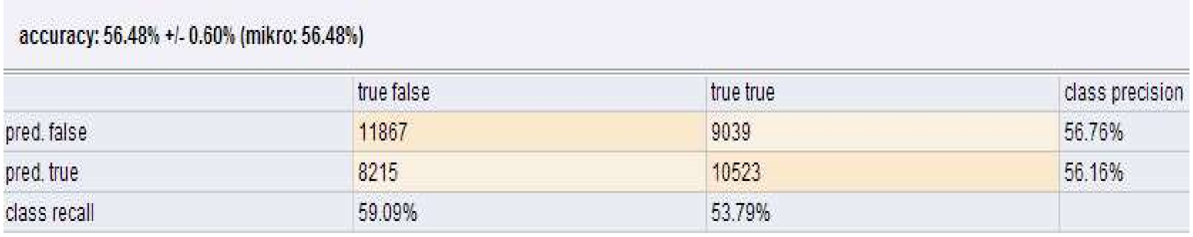
Decision Tree



accuracy: 62.87% +/- 0.55% (mikro: 62.87%)

	true false	true true	class precision
pred. false	11588	6227	65.05%
pred. true	8494	13335	61.09%
class recall	57.70%	68.17%	

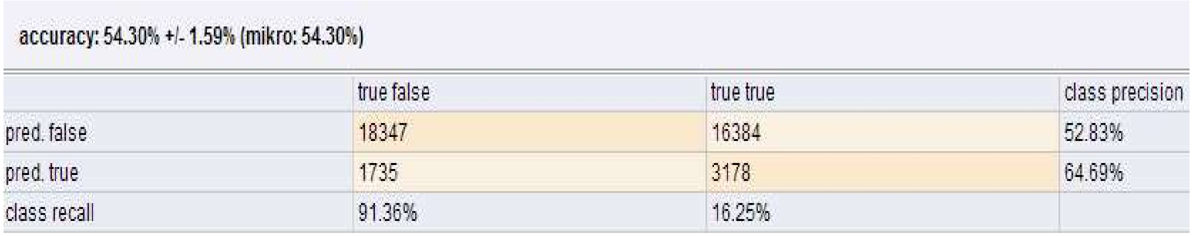
k(=5)-NN



accuracy: 56.48% +/- 0.60% (mikro: 56.48%)

	true false	true true	class precision
pred. false	11867	9039	56.76%
pred. true	8215	10523	56.16%
class recall	59.09%	53.79%	

Naïve Bayes



accuracy: 54.30% +/- 1.59% (mikro: 54.30%)

	true false	true true	class precision
pred. false	18347	16384	52.83%
pred. true	1735	3178	64.69%
class recall	91.36%	16.25%	