Classifying New Particle Formation

Group: John 117

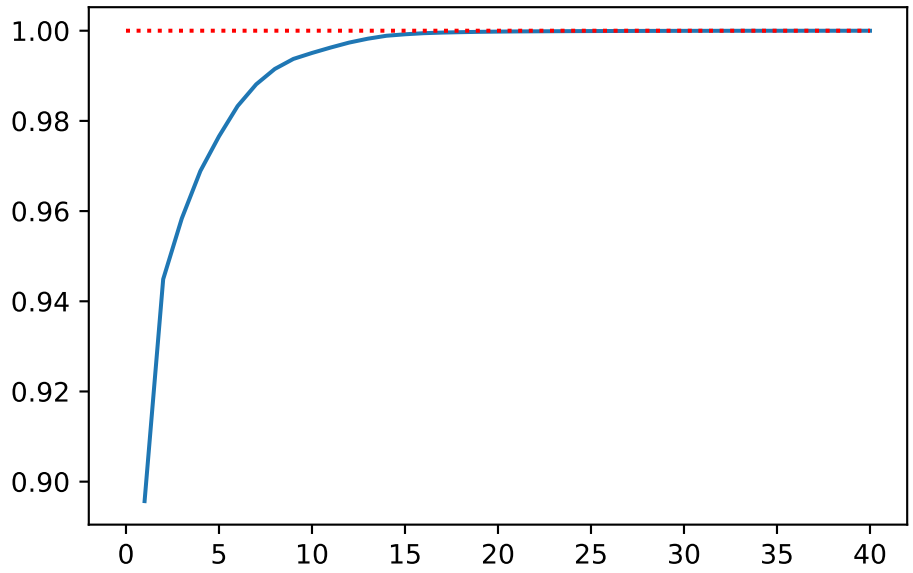
Members: Elias Toukolehto and Joacim Sarén

Preprocessing

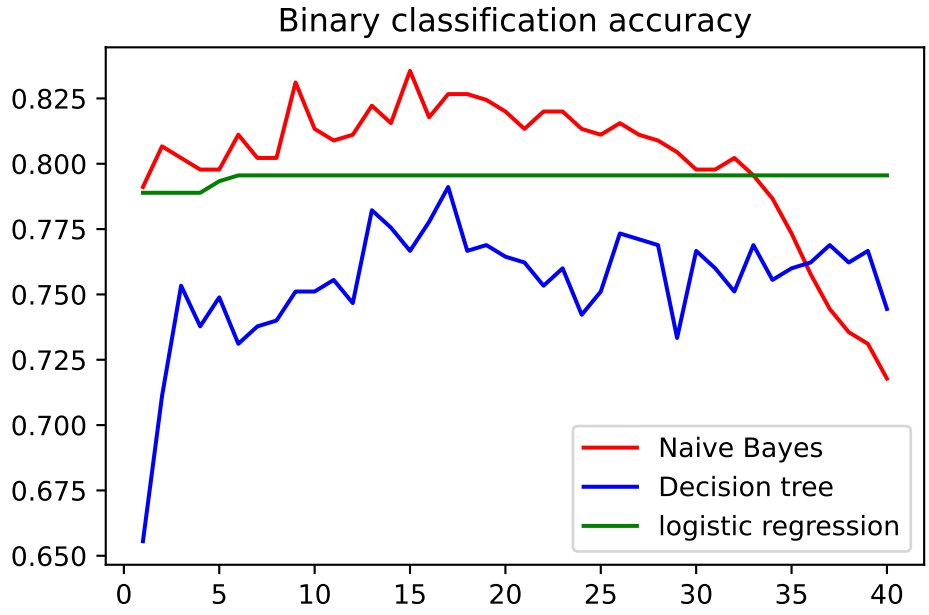
We decided to start by trying feature selection and dimensionality reduction via principal component analysis with a few different machine learning models.

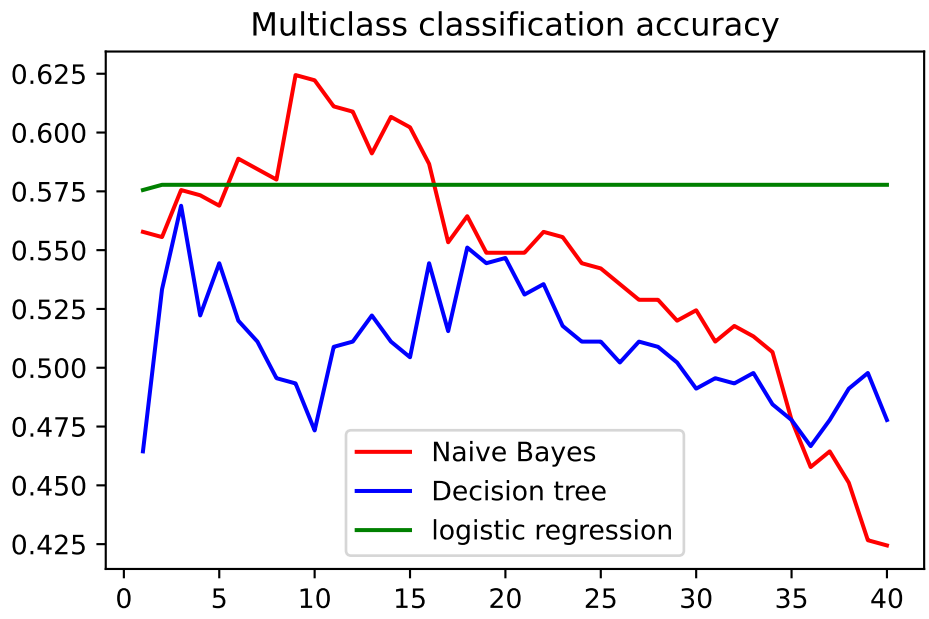
Principal component analysis

First we looked at cumulative proportion of variance explained by the principal components. We used l2 normalization. We looked at the behaviour of PCA and all the models at upto 100 components, but nothing interesting happened with any of them beyond about 40 components, so the charts here focus on 40 or less components. Accuracy is measured as the average accuracy using 10-fold cross validation.



Then we looked at the accuracy of logistic regression, decision tree and gaussian naive Bayes classifiers with different numbers of principal components for binary and multiclass classification.





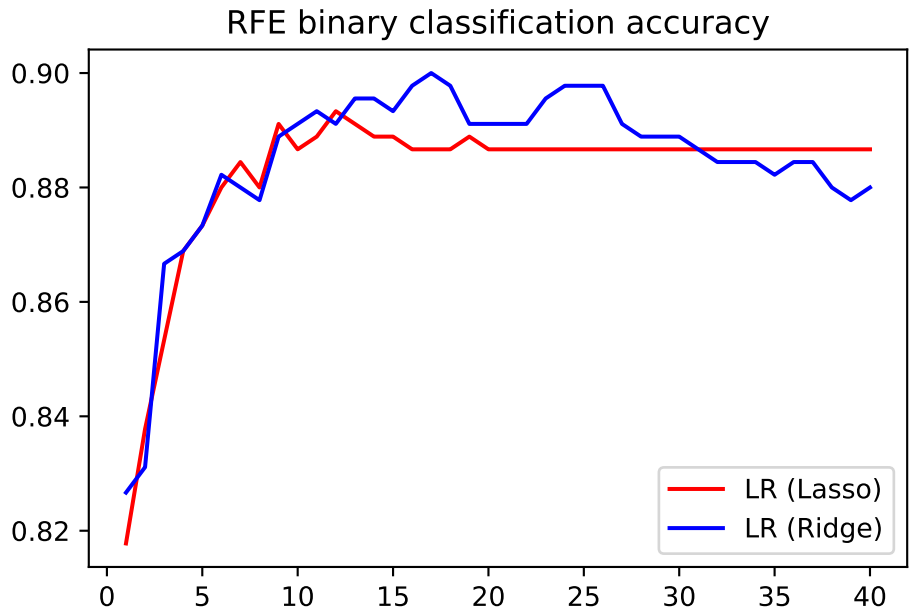
The gaussian Naive Bayes classifier with 9 predictors is clearly the best oferall performer out of the tested models.

Feature selection

Another preprocessing method is feature selection, for which we have used Recursive Feature elimination using 5-fold cross validation, which is a form of backward selection. With RFE we get the following results for binary classification

	Model	Optimal n of features	Best accuracy	Accuracy without RFE
0	LR L1	12	0.891111	0.886667
1	LR L2	17	0.897778	0.875556

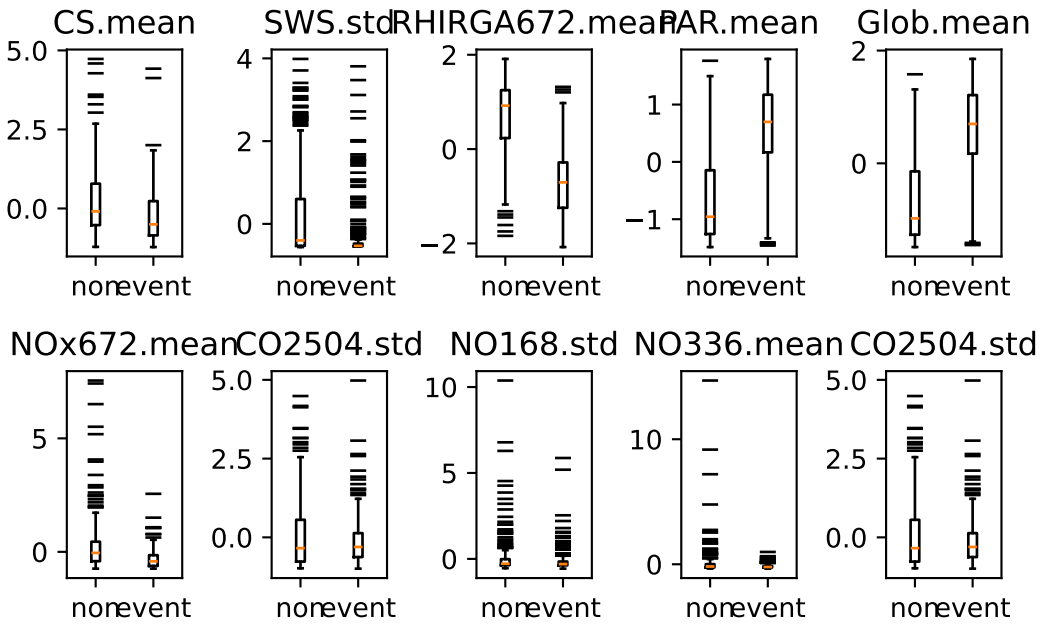
on this table we can compare the improvements RFE provides, able to increase binary classification accuracy by couple percent compared to including all features.



We can see that increasing features results in better accuracy, usually having the best accuracy at 10-20 features. We currently only have linear regression with L1 (lasso) and L2 (ridge) penalties. After comparing different models, we will probably use some combination of PCA and Feature selection in our final model

Feature analysis

Analyzing coefficients allowed us to rank some features by usefulness. The ones for the plot below were defined by performance with logistic regression using 5-fold cross validation.



Here are boxplots of some of the most useful (top row), and least useful (bottom row) features for binary classification in the dataset. For the reliable features we can see that the quartaile ranges don't have much overlap, as for the unreliable features there is a lot of overlap between classes. CS.mean is an oulkier, where models find it very useful even though it has surprisingly similar boxplots between classes.