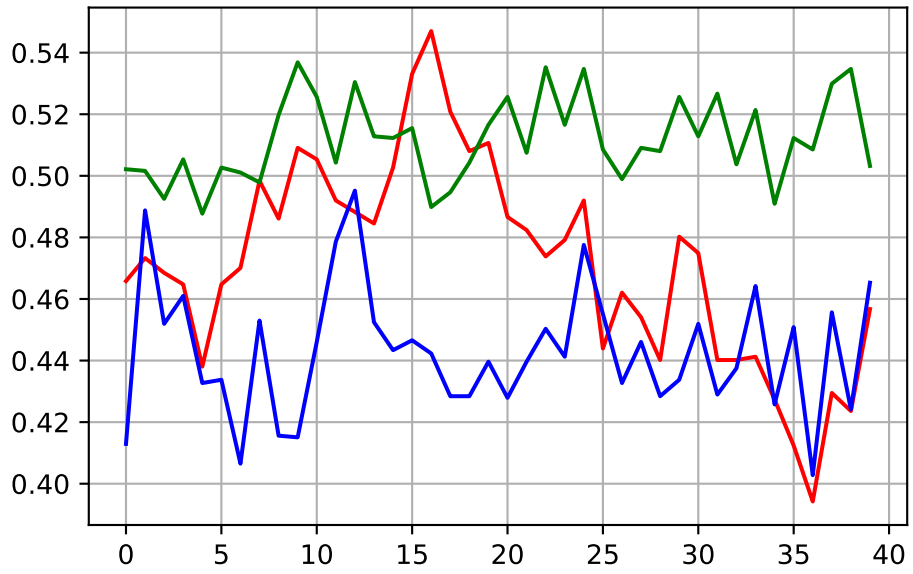
Classifying New Particle Formation



Group: John 117

Members: Elias Toukolehto and Joacim Sarén

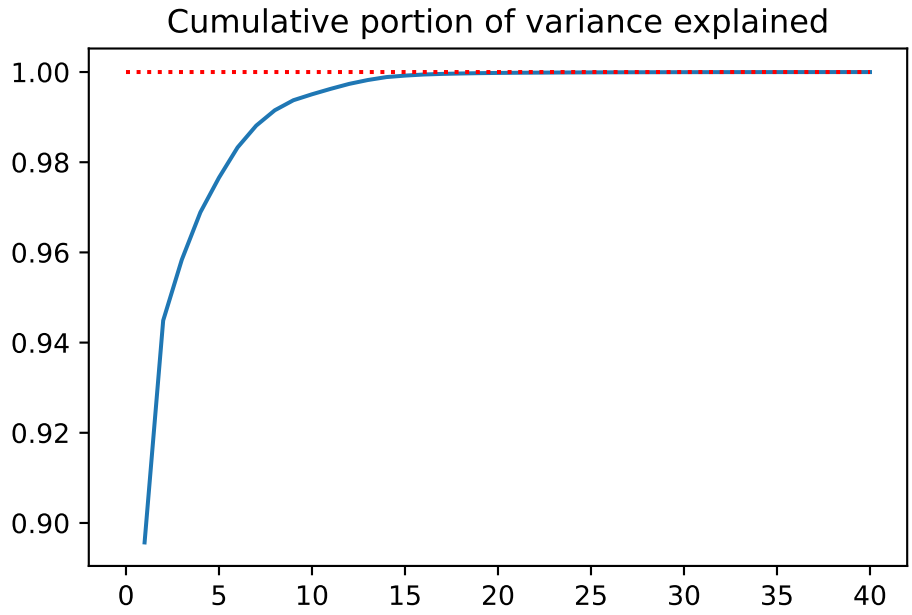
Preprocessing methods

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

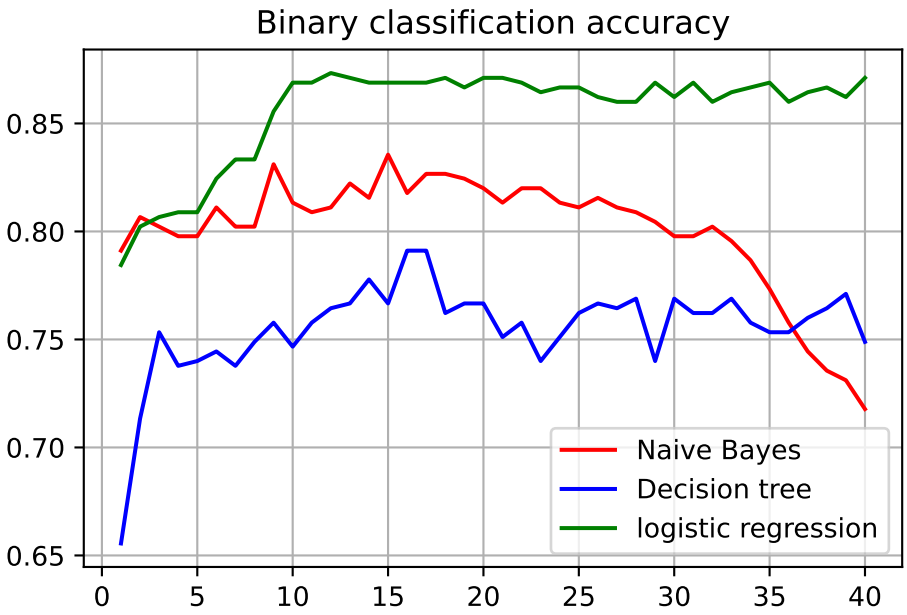
Principal component analysis

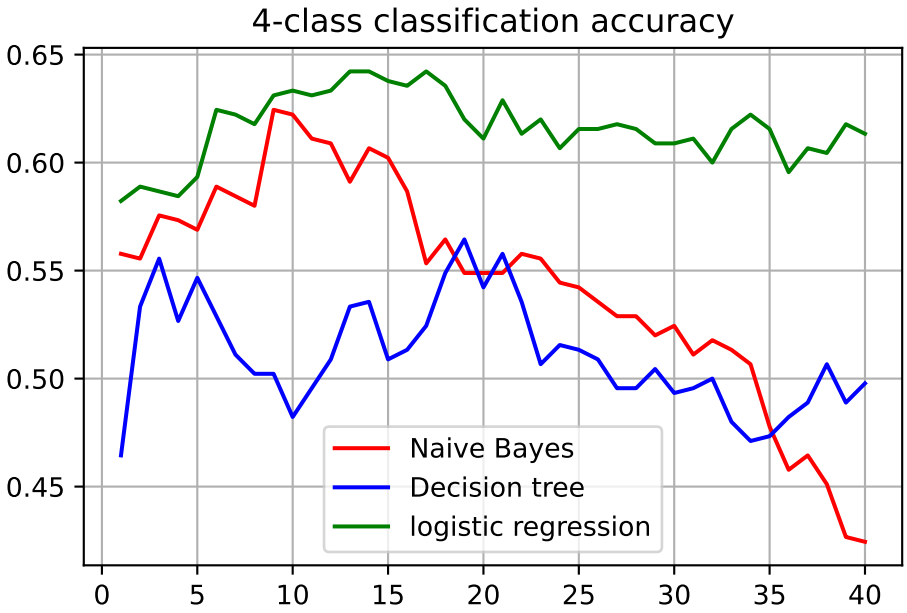
First we looked at cumulative proportion of variance explained by the principal components. We decided to use PCA for it's relavive simplicity. We used l2 normalization in order to keep

all predictors in consideration when running PCA. We looked at the behaviour of PCA and all the models at up to 100 components. We won't show data beyond 40 components, beccause it didn't show any interesting behaviour that's not visible with lower values. This is mainly to keep the charts more readable.



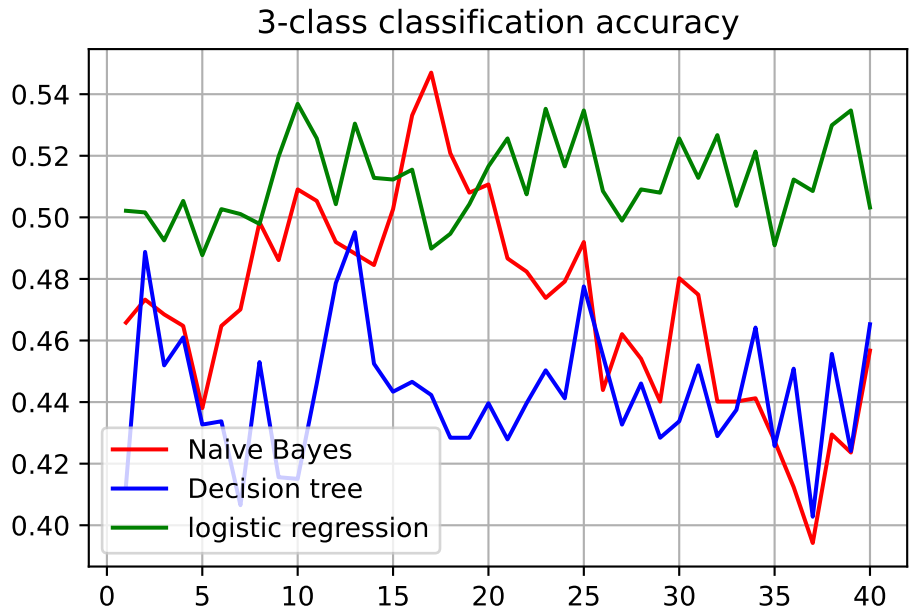
Then we looked at the accuracy of logistic regression, decision tree and gaussian naive Bayes classifiers with different numbers of principal components for binary and 4-class classification using 10-fold cross-validation.





Logistic regression achieves the highest accuracy in both binary and 4-class classification. Because the task focuses on binary classification, which is also easier than multiclass classification, we wanted to explore a 2-tier classifier. First tier seperates events and non-events, and the 2nd classifies events to the specific event classes. To choose the model for the 2nd tier, we looked at the 3-class classification accuracy of thea same classification methods trained on data only containing events.

The data with 10-fold cross-validation was inconclusive as accuracy for logistig regression jumped up and down in around the same range with 15 or less components. We suspect the main reason to be the small population of the smallest class. Out of the 225 events in the training data, only 26 are in class Ia. To smooth out the result, we decided to use 26-fold CV to match this population size. Our cross validation method maintains class sizes within folds, so the validation set always includes one Ia event.



As we hoped, there results are clearer now. Logistic regression with 22 principal components has the highest accuracy, but LR with just 9 components gets close to the same reult. Naive Bayes with 16 components is between them, and clearly the best result for NB. Because we know that each component is less important than the previous one, we decided to only test LR with 9 components and NB with 16 components in our 2-tier classifier.

PCA-based 2-tier classifier

Based on results shown in the previous chapter, we used logistic regression with 12 principal components to seperate events and non-events. For event classification we tested LR and NB with 9 and 16 components respectively.

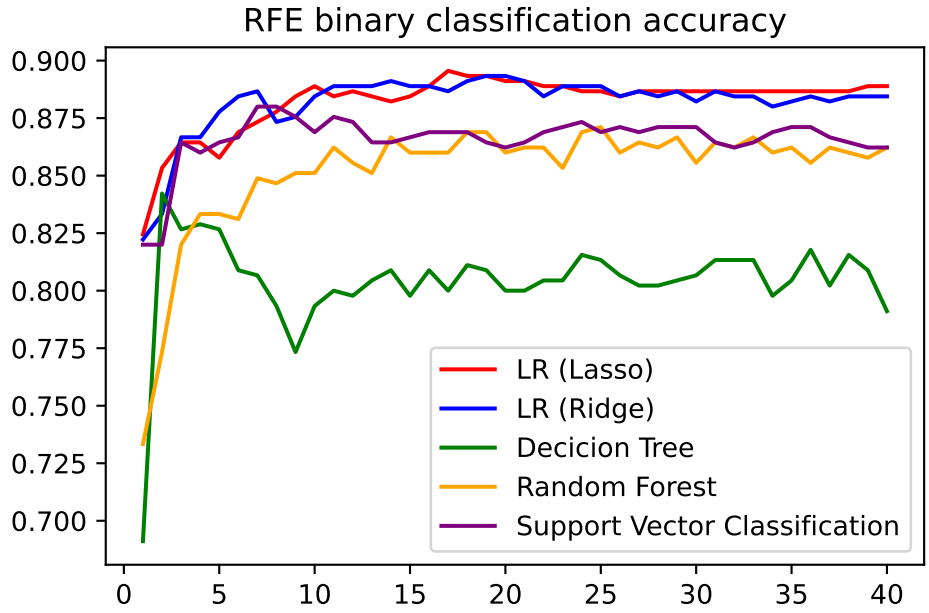
Feature selection

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	17	0.895556	0.891111
1	LR L2	19	0.893333	0.886667
2	DT	2	0.842222	0.817778
3	RF	25	0.871111	0.866667
4	SVC	7	0.880000	0.866667

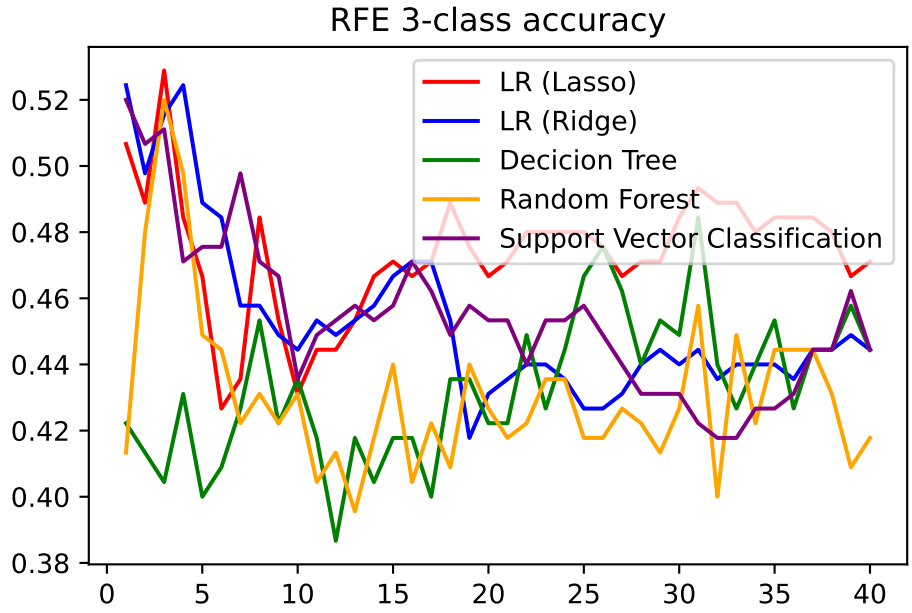
	Model	Optimal n of features	Best accuracy	Accuracy without RFE
0	LR L1	3	0.528889	0.466667
1	LR L2	1	0.524444	0.475556
2	DT	31	0.484444	0.417778
3	RF	3	0.520000	0.453333
4	SVC	1	0.520000	0.457778

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



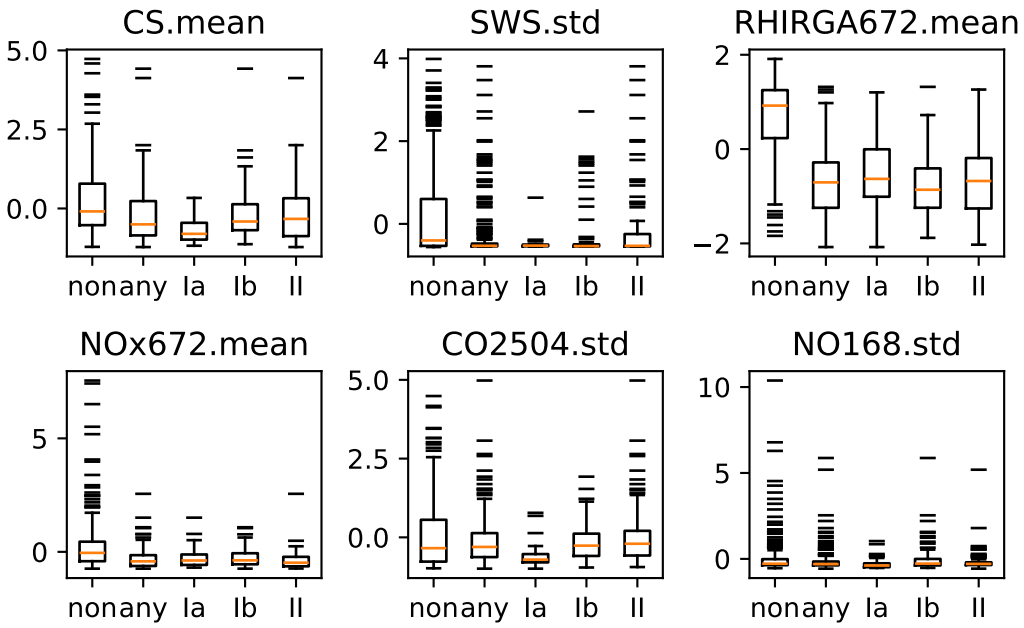
We can see that increasing features results in better accuracy, usually having the best accuracy at 10-20 features. We currently only have linear regression with L1 (lasso) and L2 (ridge)

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.