Activity prediction for chemical compounds

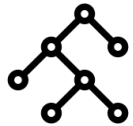
ID2214 – Data Science Project

Team members: David Happel, Albin Bååw, and Petrus Oskarsson

- **Final selected model**
- Handle of imbalance data
- Metrics
- Evaluated data sets and features
- Model comparison
- Parameter settings

The final selected model is the random forest classifier with AUC 0.77 and F1 0,69 scores on validation data

Final results



Selected model and performance

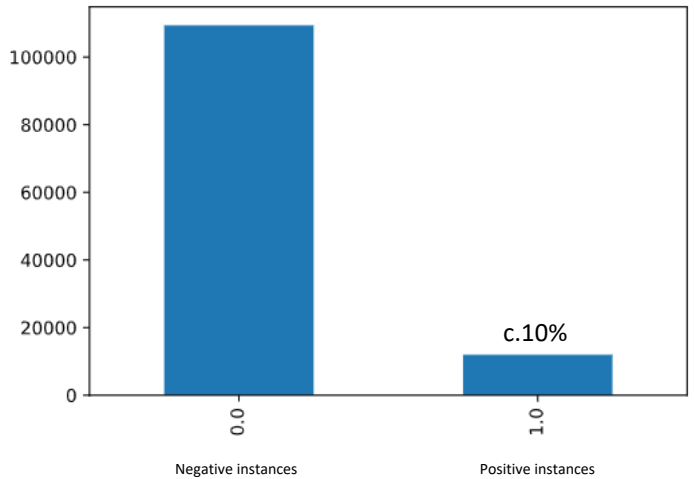
- The final selected model is the random forest classifier
- Parameter settings:
  - 291 trees of depth 8.
- Validation performance:
  - AUC: 0.7730 (+/- 0.01)
  - F1 0.6953 (+/- 0.01)

- Final selected model
- **Handle of imbalance data**
- Metrics
- Evaluated data sets and features
- Model comparison
- Parameter settings

Initial observation: c. 10 % of positive cases – undersampling have been made to balance it out

Handle of imbalanced data set

Imbalanced data set



Actions

- The imbalance could cause potential learning issues, e.g. learning to predict all instances and get and get a 90% acc.
- Undersampling has been applied to balance out the positive and negative cases
  - Both over and undersampling was evaluated – undersampling gave the best performance

- Final selected model
- Handle of imbalance data
- **Metrics**
- Evaluated data sets and features
- Model comparison
- Parameter settings

Multiple metrics have been applied, however, F1 and AUC score are thought to be the most important

Metrics

Metrics

- Accuracy – can be misleading…
- **AUC** – Performance for every threshold + competition metrics…
- Precision – How precise of classified positives
- Recall – Fraction of actual positives
- **F1 score** - Balances precision and recall

- Final selected model
- Handle of imbalance data
- Metrics
- **Evaluated data sets and features**
- Model comparison
- Parameter settings

Seven alternative data sets have been evaluated

Evaluated data sets and features

Data sets

1. All features and **minmax** scaling
2. All features and **zscore** scaling
3. All features **except 'Lipinski'**
4. Only features **['nrAtoms', 'ExactMolWT', 'Fragments']**
5. Only features **['nrAtoms', 'ExactMolWT', 'Fragments', 'Lipinski']**
6. **Only fingerprint binaries**
7. **PCA** dimensionality reduction

Features

- 'nrAtoms'
- 'ExactMolWT'
- 'Fragments'
- 'Lipinski'
- 124 bits fingerprint vector

- Final selected model
- Handle of imbalance data
- Metrics
- Evaluated data sets and features
- **Data set and model comparison**
- Parameter settings

Best performance: data set 3 (zscore and no ‘Lipinski’) and the random forest classifier model

Data set and model comparison

Selected data sets and model comparison

Best performer: zscore and ’Lipinski’ dropped

model	AUC score	f1 score
logisticregression	0.6574 (+/- 0.00)	0.6111 (+/- 0.00)
randomforestclassifier	0.7558 (+/- 0.01)	0.6794 (+/- 0.01)
gaussiannb	0.6310 (+/- 0.01)	0.6415 (+/- 0.01)
adaboostclassifier	0.5761 (+/- 0.01)	0.5762 (+/- 0.01)
mlpclassifier	0.6877 (+/- 0.00)	0.6352 (+/- 0.01)
gradientboostingclassifier	0.6870 (+/- 0.00)	0.6294 (+/- 0.00)

Table 3: 5-fold CV Scores Dataset 3

Only fingerprint showed **similar performance**

model	AUC score	f1 score
logisticregression	0.6476 (+/- 0.01)	0.6041 (+/- 0.01)
randomforestclassifier	0.7519 (+/- 0.01)	0.6735 (+/- 0.01)
gaussiannb	0.6232 (+/- 0.01)	0.6300 (+/- 0.01)
adaboostclassifier	0.5788 (+/- 0.01)	0.5836 (+/- 0.01)
mlpclassifier	0.6775 (+/- 0.01)	0.6340 (+/- 0.00)
gradientboostingclassifier	0.6802 (+/- 0.01)	0.6184 (+/- 0.01)

Table 6: 5-fold CV Scores Dataset 6

Excluding figerprint **decreased performance**…

model	AUC score	f1 score
logisticregression	0.5596 (+/- 0.00)	0.5371 (+/- 0.01)
randomforestclassifier	0.5529 (+/- 0.01)	0.5426 (+/- 0.01)
gaussiannb	0.5429 (+/- 0.01)	0.6369 (+/- 0.00)
adaboostclassifier	0.5501 (+/- 0.01)	0.5498 (+/- 0.01)
mlpclassifier	0.5682 (+/- 0.00)	0.5651 (+/- 0.02)
gradientboostingclassifier	0.5662 (+/- 0.01)	0.5709 (+/- 0.02)

Table 5: 5-fold CV Scores Dataset 5

PCA 2 dimension – **not helpful**

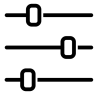
model	AUC score	f1 score
logisticregression	0.5551 (+/- 0.00)	0.5337 (+/- 0.01)
randomforestclassifier	0.5224 (+/- 0.01)	0.5157 (+/- 0.01)
gaussiannb	0.5570 (+/- 0.00)	0.5376 (+/- 0.01)
adaboostclassifier	0.5050 (+/- 0.01)	0.5070 (+/- 0.01)
mlpclassifier	0.5615 (+/- 0.01)	0.5295 (+/- 0.02)
gradientboostingclassifier	0.5597 (+/- 0.01)	0.5479 (+/- 0.01)

Table 7: 5-fold CV Scores Dataset 7

- Final selected model
- Handle of imbalance data
- Metrics
- Evaluated data sets and features
- Model comparison
- **Parameter settings**

Hyperparameter tuning of RF performed with AUC 0.77 and F1 0,69 on the validation data

Hypertuning



Parameter settings and performance

- Parameter settings:
  - 400 trees of depth None
- Validation performance:
  - AUC: 0.7730 (+/- 0.01)
  - F1 0.6953 (+/- 0.01)

rank	mean f1 score	no. trees	depth
1	0.6953 (+/- 0.01)	300	None
2	0.6886 (+/- 0.00)	200	None
3	0.6859 (+/- 0.01)	200	15
4	0.6854 (+/- 0.01)	300	15
5	0.6810 (+/- 0.01)	100	None

Table 8: Top 5 best performing configurations based on f1 score

rank	mean AUC score	no.trees	depth
1	0.7730 (+/- 0.01)	300	None
2	0.7700 (+/- 0.01)	200	None
3	0.7628 (+/- 0.01)	300	15
4	0.7582 (+/- 0.01)	200	15
5	0.7547 (+/- 0.01)	100	None

Table 9: Top 5 best performing configurations based on AUC score

Questions?

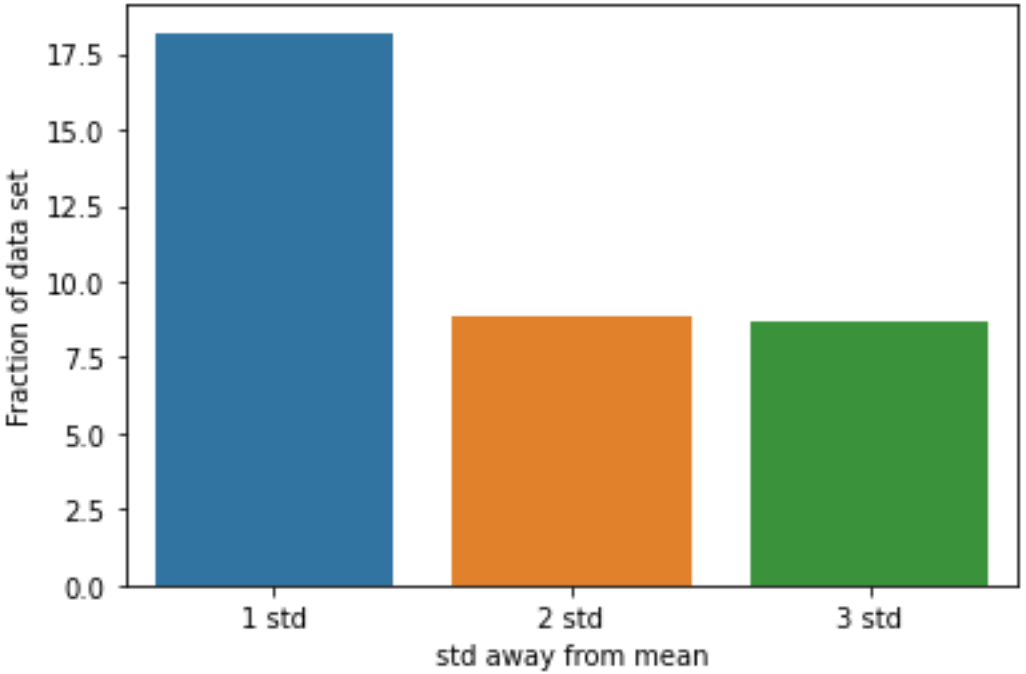
Team members: David Happel, Albin Bååw, and Petrus Oskarsson

Appendix A – Data Fram view

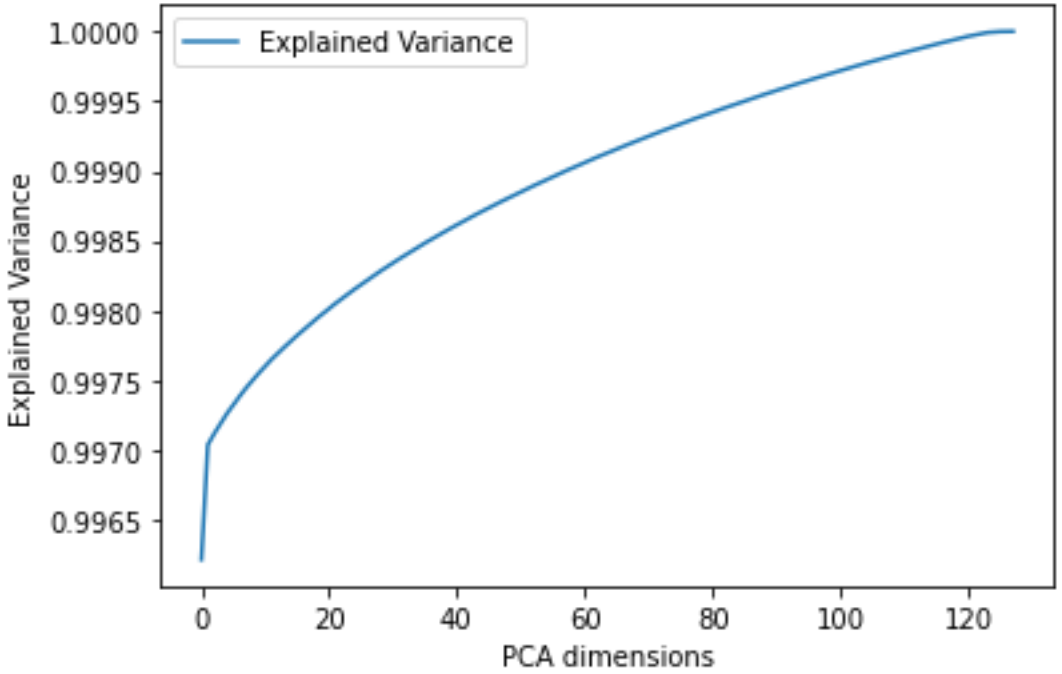
	INDEX	SMILES	nrAtoms	ExactMolWT	Fragments	Lipinski	fp_0	fp_1	fp_2	fp_3	...	fp_114	fp_115	fp_116	fp_117	fp_118	fp_119	fp_120	fp_1
0	121375.0	0.0	28.0	390.115047	0.0	28.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0
1	121376.0	0.0	28.0	381.185255	0.0	28.0	0.0	0.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1
2	121377.0	0.0	20.0	282.096420	0.0	20.0	0.0	0.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	121378.0	0.0	26.0	391.029663	0.0	26.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0
4	121379.0	0.0	34.0	500.191583	0.0	34.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
40453	161828.0	0.0	25.0	352.099397	0.0	25.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0
40454	161829.0	0.0	26.0	416.037975	0.0	26.0	0.0	0.0	1.0	0.0	...	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1
40455	161830.0	0.0	26.0	353.101171	0.0	26.0	0.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0
40456	161831.0	0.0	25.0	338.137890	0.0	25.0	1.0	1.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
40457	161832.0	0.0	31.0	447.102846	0.0	31.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1

40458 rows × 130 columns

Appendix B - outliers



Appendix C – PCA explained variance



Appendix D – Hierarchical clustering