ID 2214 MOLECULE ACTIVITY CLASSIFIER

LUCAS LARSSON & MIHAELA BAKŠIĆ

STEPS

01

FEATURE SELECTION

Start by researching and defining features

02

TRAINING PROCESS

Train the model using the previous selected features

03

RESULTS VALIDATION

Examine the results

04

REPEAT

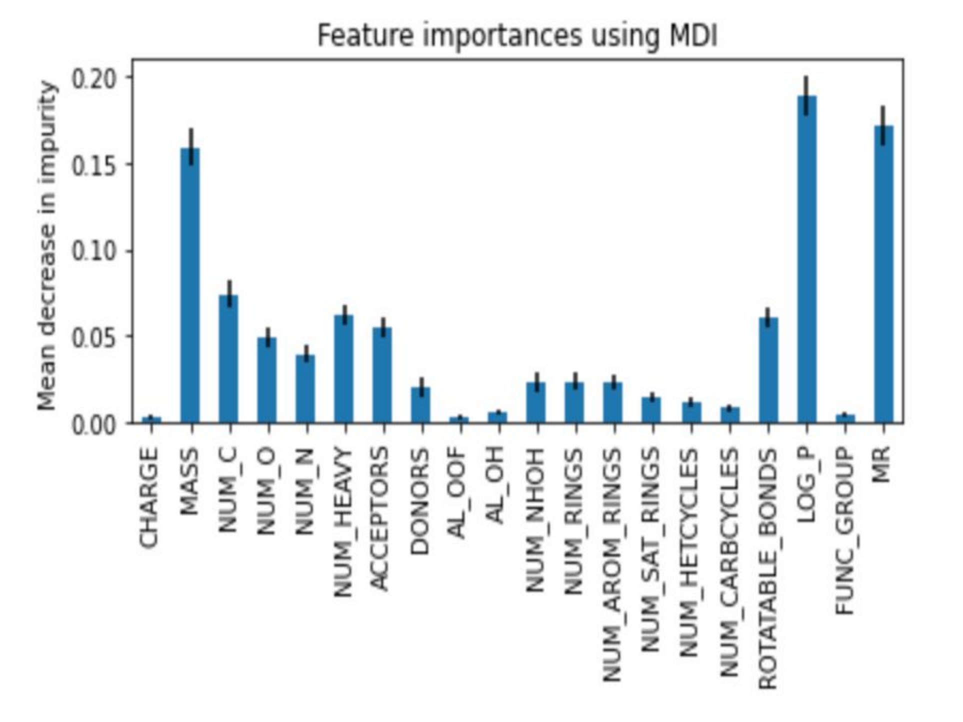
Repeat till perfect

FEATURE SELECTION

FEATURE SELECTION

FEATURE IMPORTANCE WITH MDI

Optimization was carried out by removing features with MDI below a certain threshold while observing change in AUC



FEATURES	AUC
MASS, NUM C, NUM O, NUM N, NUM HEAVY, ACCEPTORS, DONORS, NUM NHOH, NUM RINGS, NUM AROM RINGS, NUM SAT RINGS	0.9492
MASS, NUM C, NUM O, NUM N, NUM HEAVY, ACCEPTORS, DONORS, NUM NHOH, NUM RINGS, NUM AROM RINGS, ROTATABLE BONDS, LOG P, MR	0.9464
MASS, NUM C, NUM HEAVY, ACCEPTORS, ROTATABLE BONDS, LOG P, MR	0.9122
MASS, LOG P, MR	0.8107

FEATURE SELECTION

Morgan Fingerprints

Fingerprints extracted using RDKit, Multiple lengths were tested in the span of 10–1024.
Optimal length was 512 bits.

WHEN AND WHY?

SMOTE

OVERSAMPLING

ROS

COMPARISON

	SMOTE	ROS
COMPUTE	Expensive	Efficient
REDUCE OVERFITTING	Interpol existing observations	Repeats existing observations
NUMBER OF REPEATED OBSERVATIONS	Not Sensitive	Sensitive

MODELS

MODELS

Logistic regression

Naive Bayes

Random Forest

XGBoost

RESULTS

RESULTS

	Random Forest
AUC	0.938
ACC	0.988
F1	0.929

FP With SMOTE

Do you have any questions?

Lucas & Mihaela

THANKS

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik