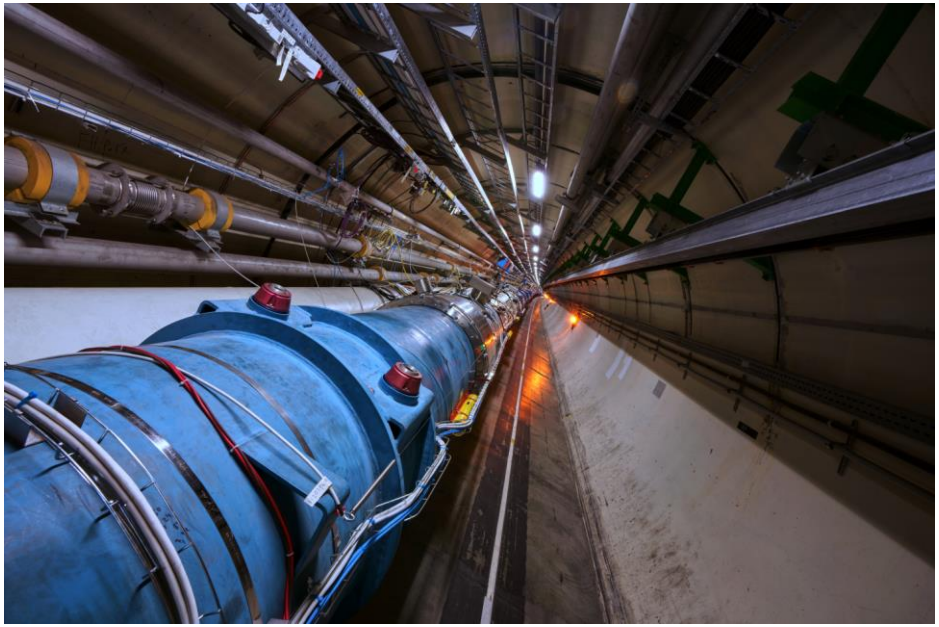
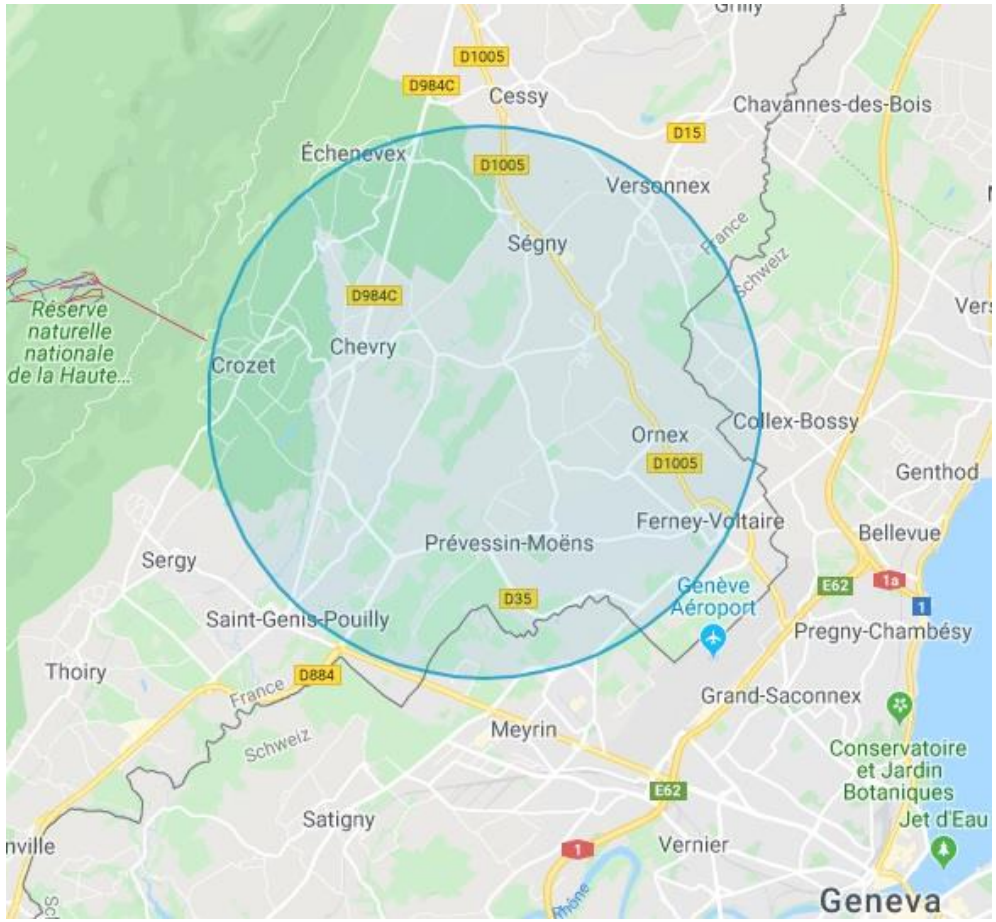
Development of Tools for Data Quality Control for ALICE Experiment at CERN, Using Machine Learning Methods

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY

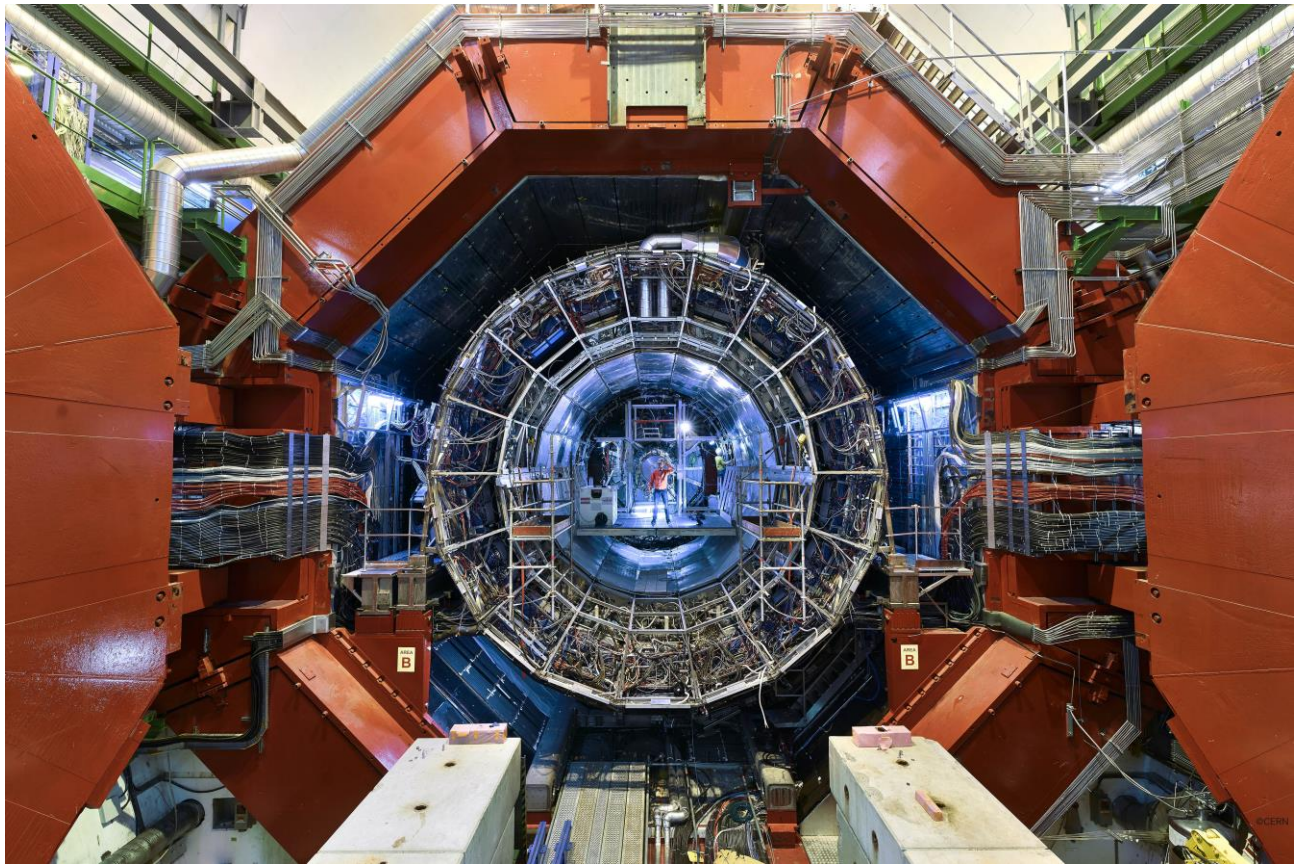
Bartłomiej Cerek

CERN




D1005
D984C
Cessy
Chavannes-des-Bois
D15
Échenevex
D1005
Versonnex
Ségny
France
Schweiz
D984C
Réserve naturelle nationale de la Haute...
Chevry
Crozet
Collex-Bossy
Ornex
D1005
Genthod
Ferney-Voltaire
Bellevue
Prévessin-Moëns
Sergy
Genève Aéroport
E62
1a
D35
Saint-Genis-Pouilly
Pregny-Chambésy
Thoiry
D884
France
Grand-Saconnex
Schweiz
Meyrin
Conservatoire et Jardin Botaniques
E62
Satigny
Jet d'Eau
Vernier
1
Geneva

ALICE



Heavy ions (like Pb)

Quantum chromodynamics

10 000 tones

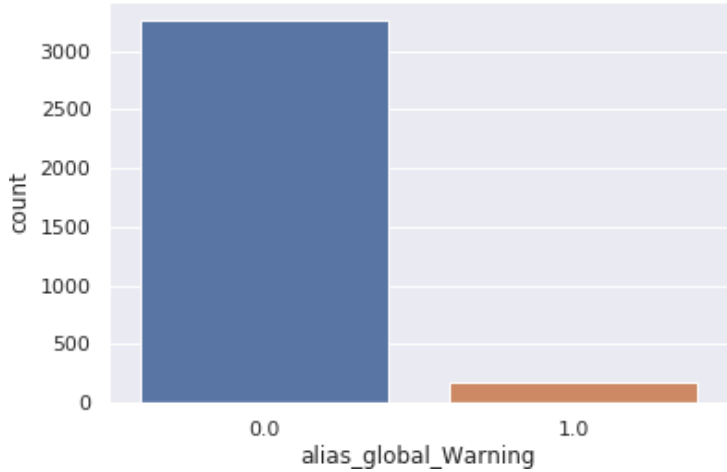
After 2021 - 3 TB/s

Anomaly Detection

run	chunkID	time	year	period.fString	pass.fString	dataType.fString	startTimeGRP	stopTimeGRP	duration
287000	1	1527204124	2018	LHC18f	pass1	NaN	1527204124	1527231797	27673
287000	2	1527204124	2018	LHC18f	pass1	NaN	1527204124	1527231797	27673
287000	3	1527204124	2018	LHC18f	pass1	NaN	1527204124	1527231797	27673
287000	4	1527204124	2018	LHC18f	pass1	NaN	1527204124	1527231797	27673
287000	5	1527204124	2018	LHC18f	pass1	NaN	1527204124	1527231797	27673

RUN ~ 12h
CHUNK ~ 8–15 min

SAMPLES - 3429
FEATURES - 231



Anomaly Detection

	bz	meanTPCnclF	meanTPCChi2	rmsTPCChi2	slopeATPCnclF	slopeCTPCnclF	slopeATPCnclFErr	slopeCTPCnclFErr
mean	0.671010	0.639430	0.418191	0.354881	0.403929	0.617189	0.039607	0.617189
std	0.469882	0.082914	0.107004	0.078426	0.116310	0.094160	0.046425	0.094160

2 rows × 133 columns

	bz	meanTPCnclF	meanTPCChi2	rmsTPCChi2	slopeATPCnclF	slopeCTPCnclF	slopeATPCnclFErr	slopeCTPCnclFErr
0	0.000000	0.650378	0.586648	0.299807	0.567304	0.506033	0.029799	0.506033
1	0.000000	0.652363	0.579001	0.299619	0.553643	0.490807	0.028887	0.490807
2	0.000000	0.656779	0.582115	0.302841	0.576802	0.486826	0.028507	0.486826
54	0.000000	0.684929	0.573975	0.319703	0.633521	0.722099	0.480131	0.722099
69	0.000000	0.676676	0.583633	0.238868	0.657002	0.320537	0.628478	0.320537
182	0.000000	0.686070	0.549347	0.261746	0.642229	0.560798	0.228963	0.560798

Technology


python™


jupyter
CO


NumPy


scikit
learn

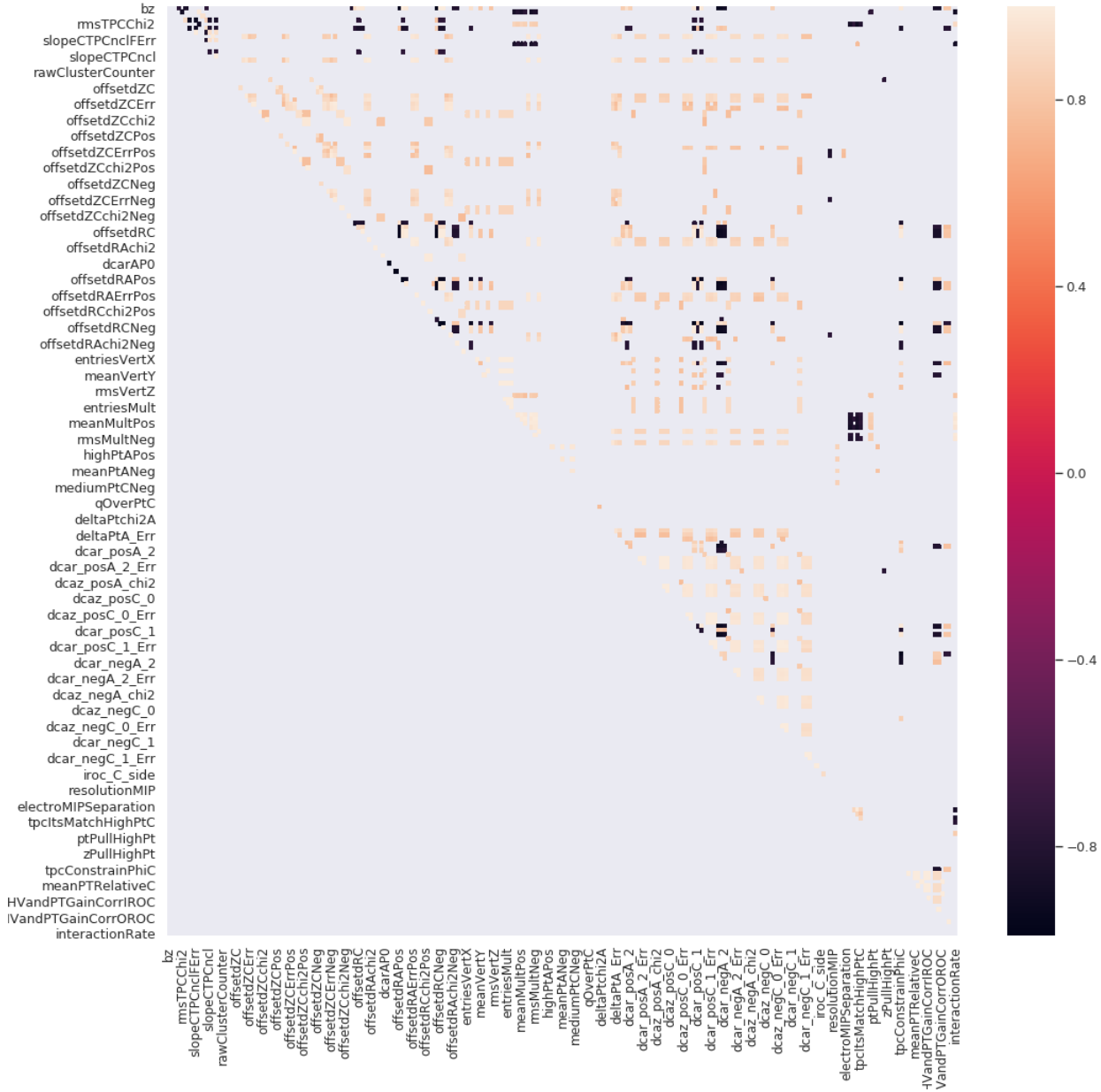

pandas


PyTorch


matplotlib

Data Analysis

Features nessesary to mimic results of former agorithm with logistic regression:



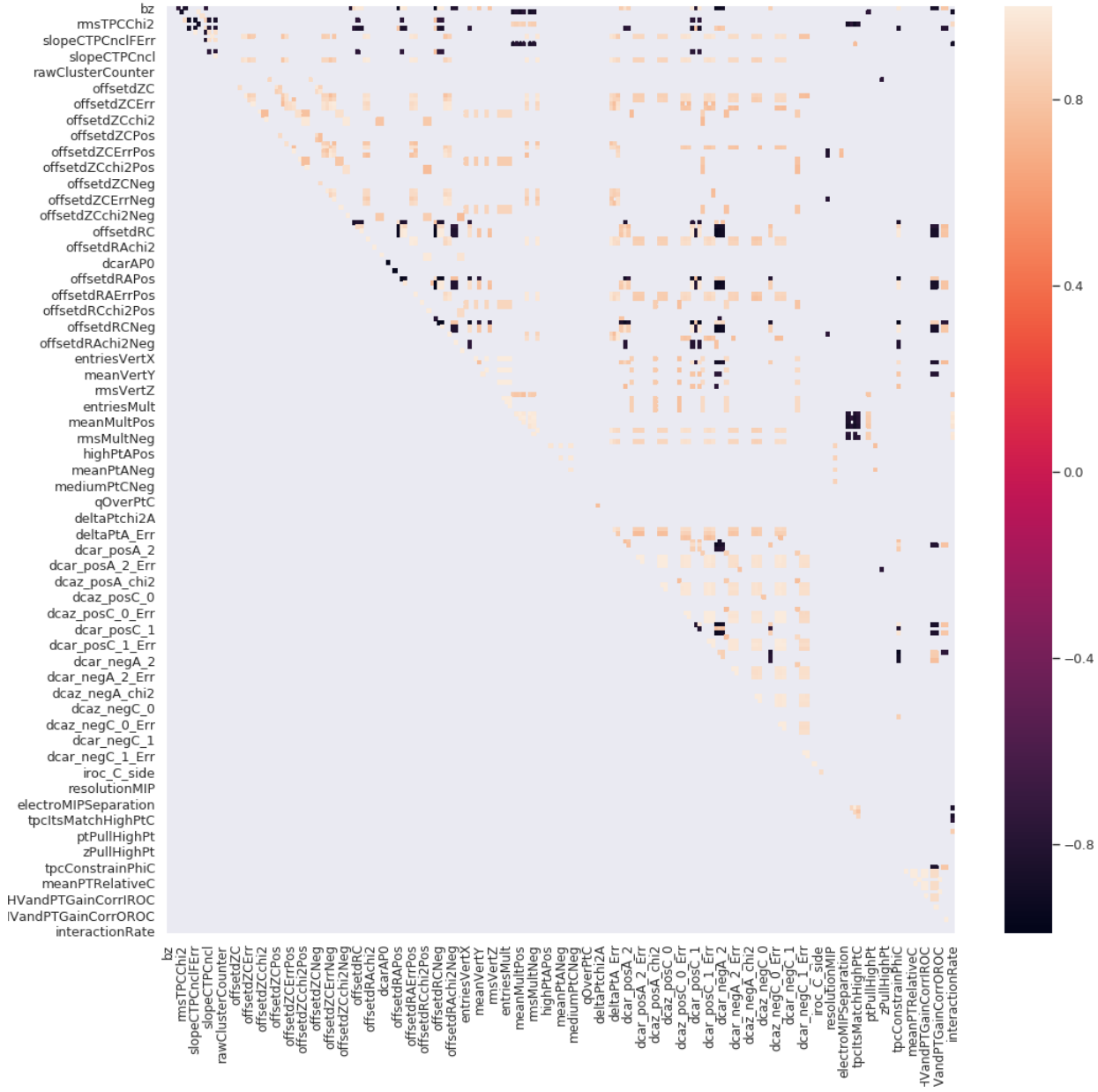
Data Analysis

Features nessesary to mimic results of former agorithm with logistic regression:

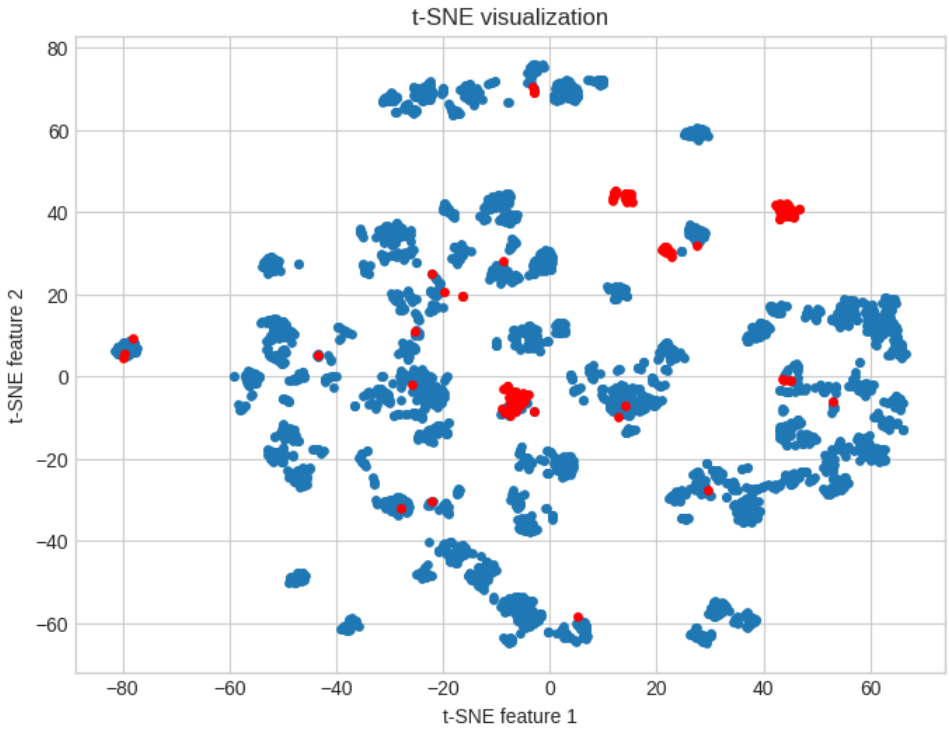
96% accuracy / 93% b.accuracy

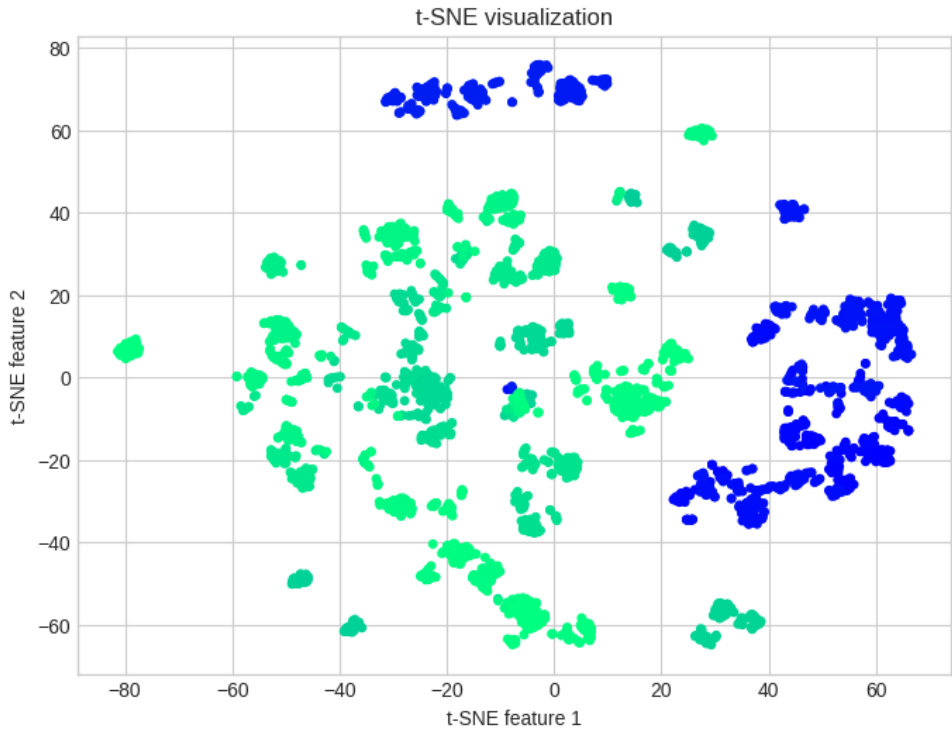
~40 / 231 **???**

Lightweight Dataset: **133**



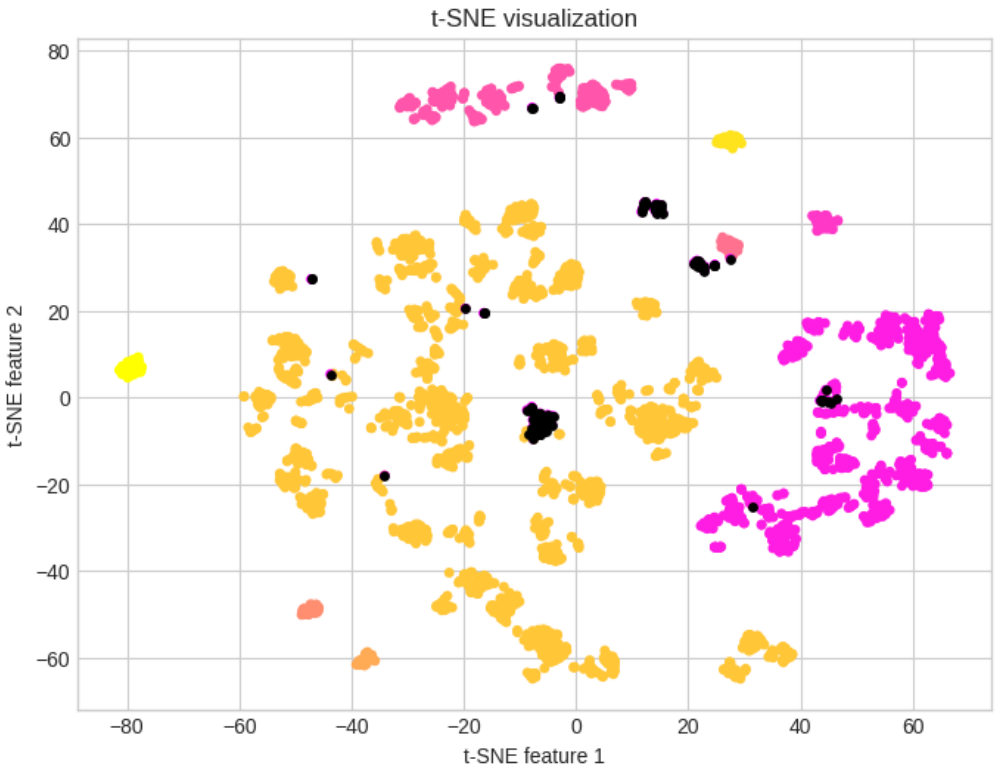
Classic Machine Learning



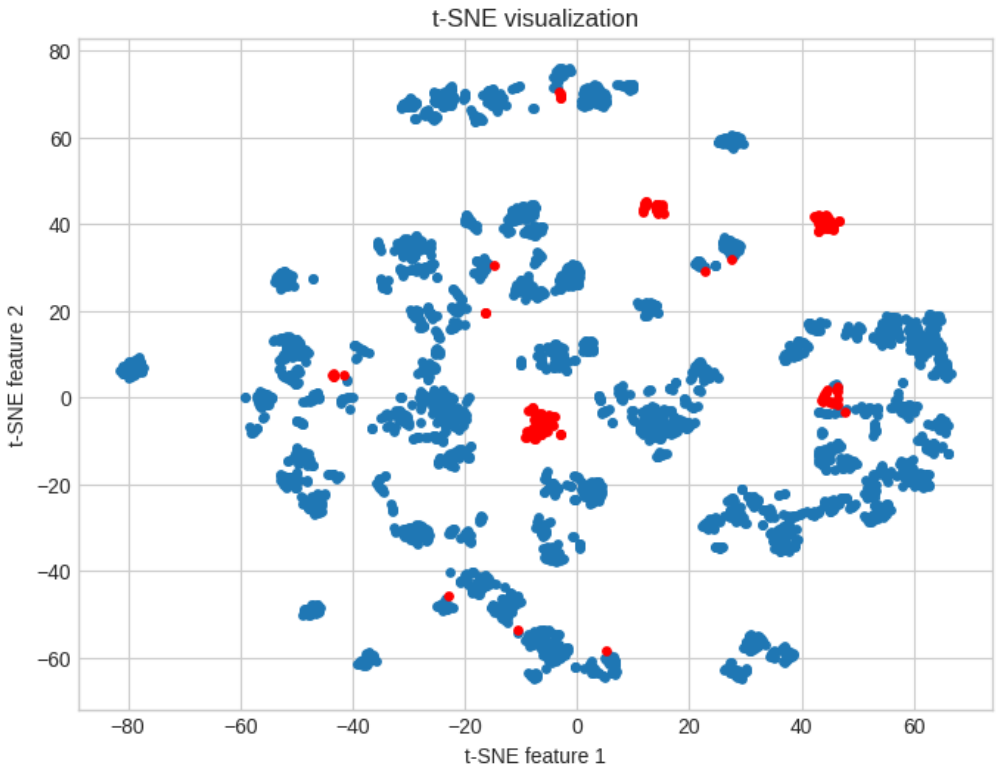


Classic Machine Learning

DBSCAN

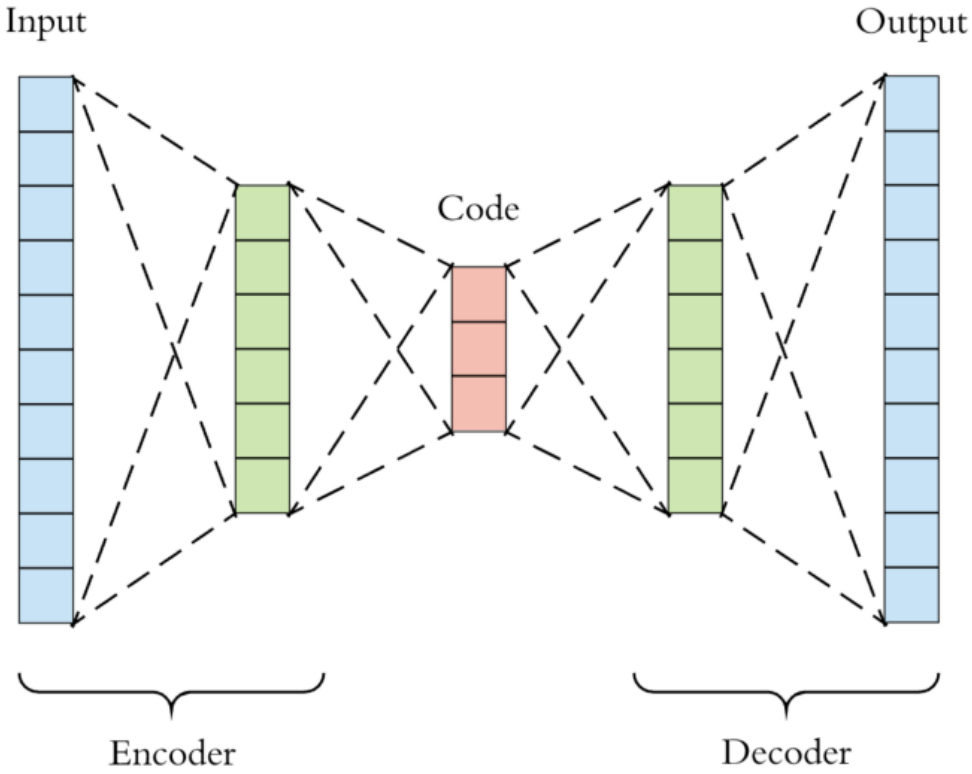


Isolation Forest

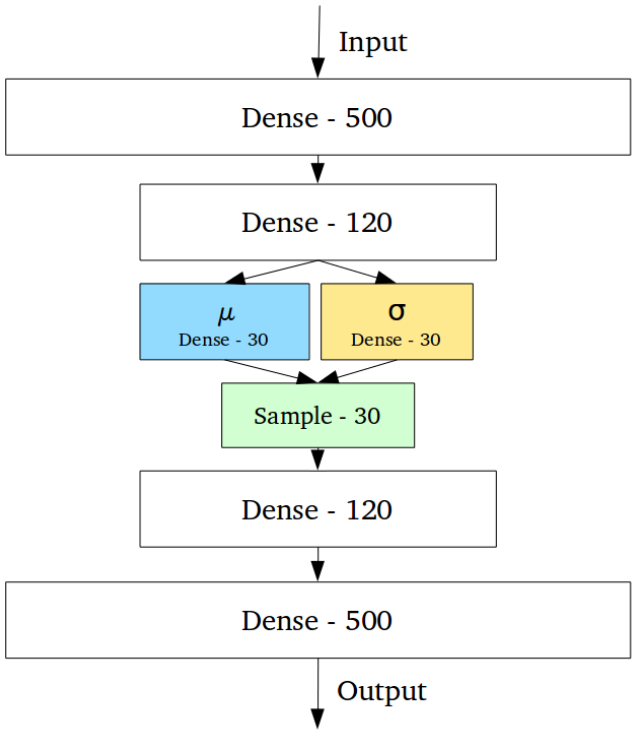


Deep Autoencoders

Simple

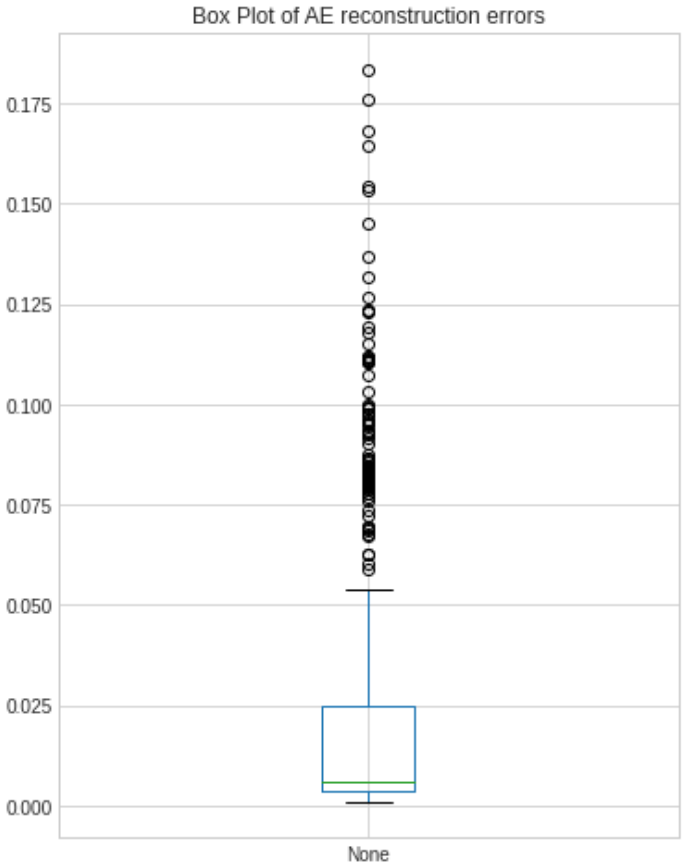


Variational

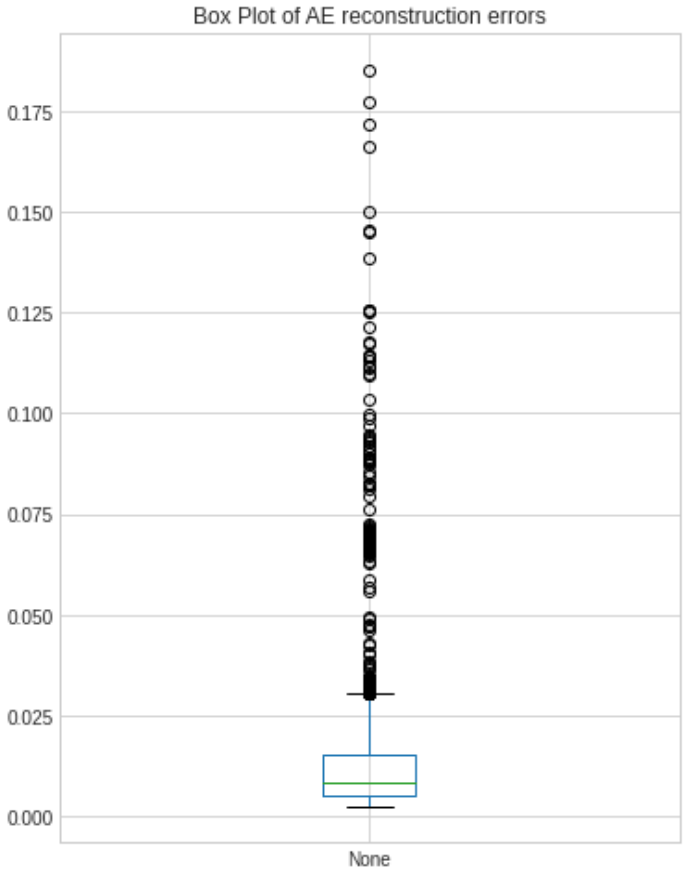


Deep Autoencoders

Simple

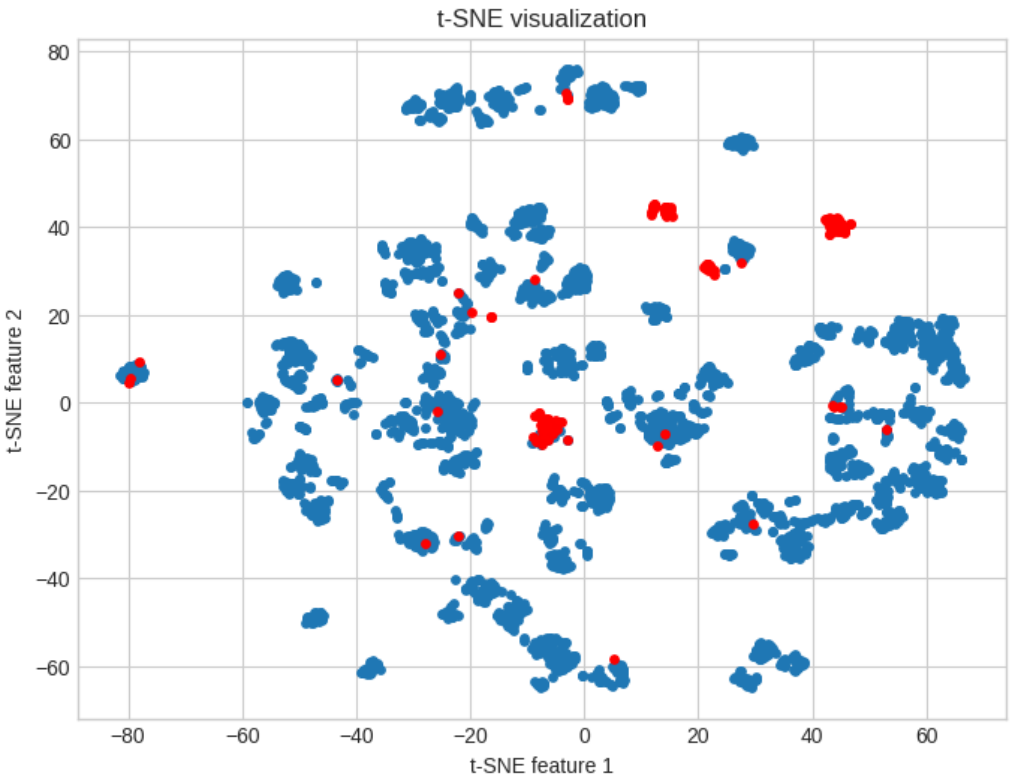


Variational

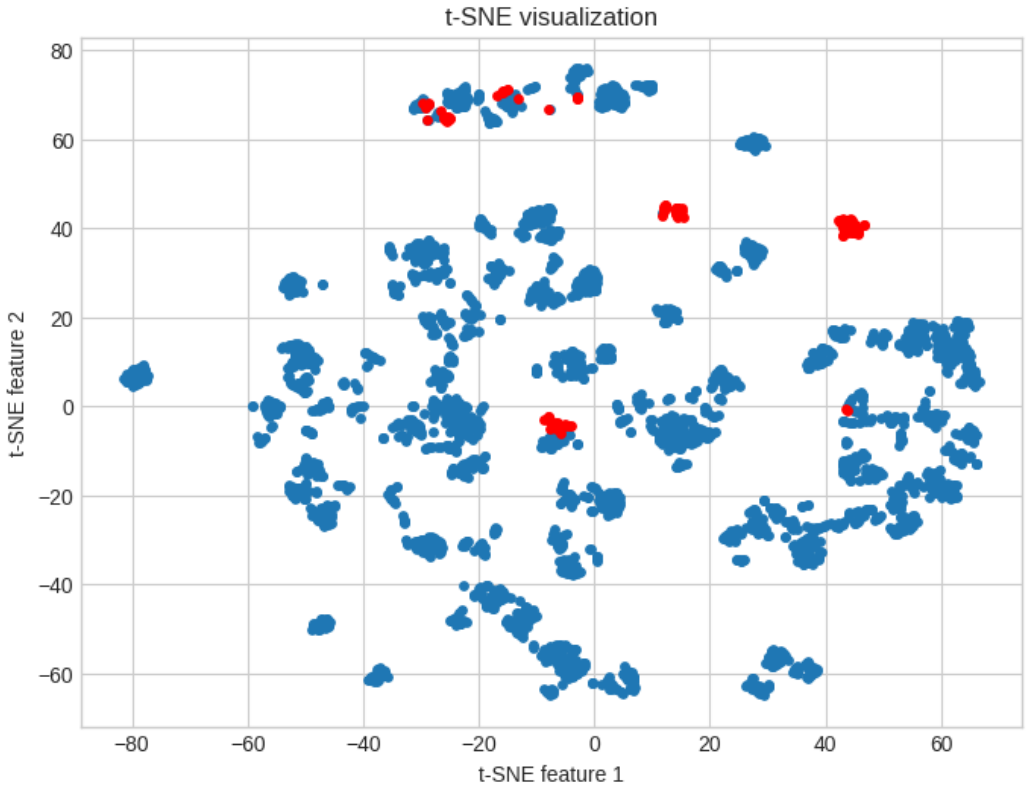


Deep Autoencoders

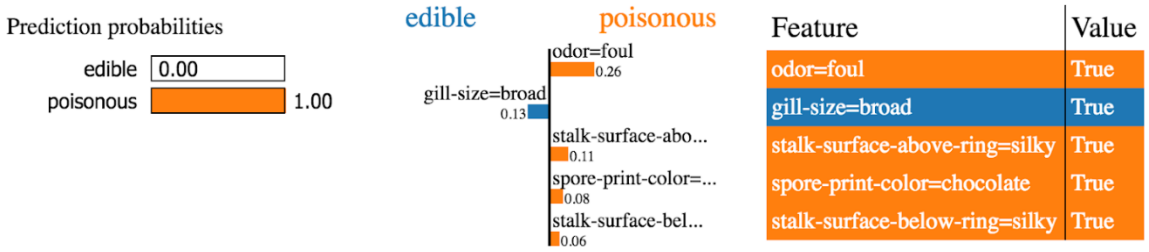
Former



Simple Autoencoder



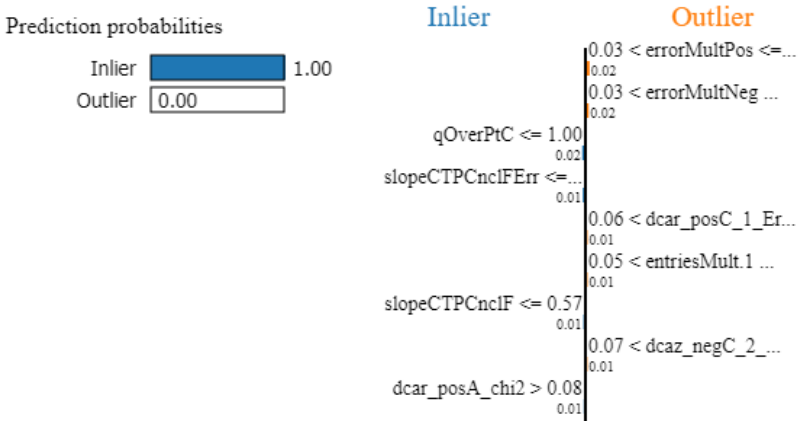
Model explainability / LIME



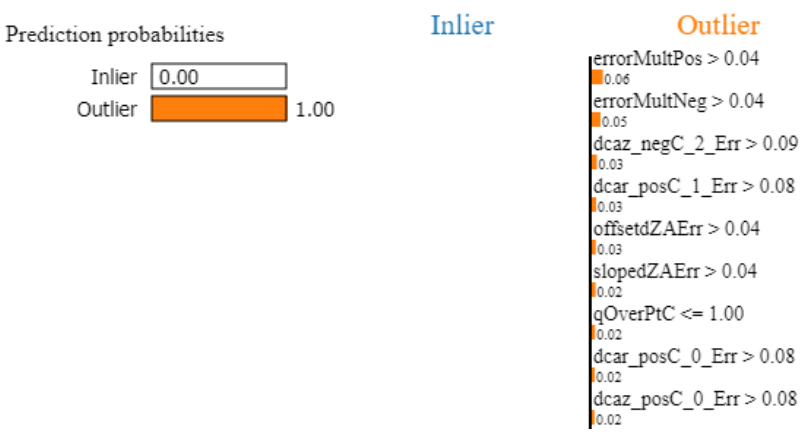
What about unsupervised cases?

RAW DATA → (Isolation Forest assigns labels) → LABELD DATA → (Random Forest copies iForest) → TRAINED MODEL → (LIME explains classifications) → EXPLAN ATION

LIME with Isolation Forest

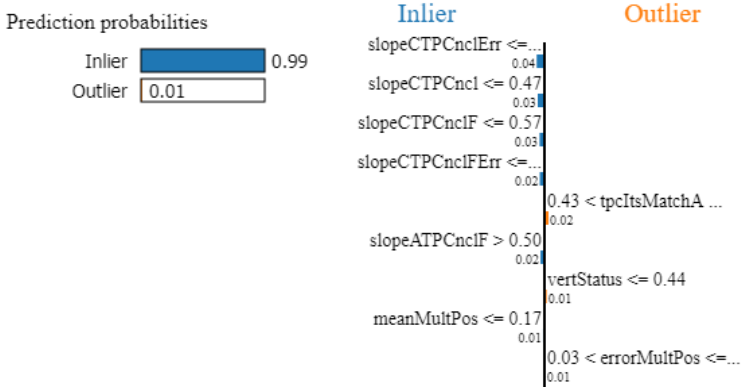


Feature	Value
errorMultPos	0.03
errorMultNeg	0.03
qOverPtC	1.00
slopeCTPCnclFErr	0.51
dcar_posC_1_Err	0.06
entriesMult.1	0.07
slopeCTPCnclF	0.51
dcaz_negC_2_Err	0.07
dcar_posA_chi2	0.10
slopeCTPCnclErr	0.43
dcaz_posC_0_Err	0.06
dcar_posC_0_Err	0.06
dcaz_posC_1_Err	0.06

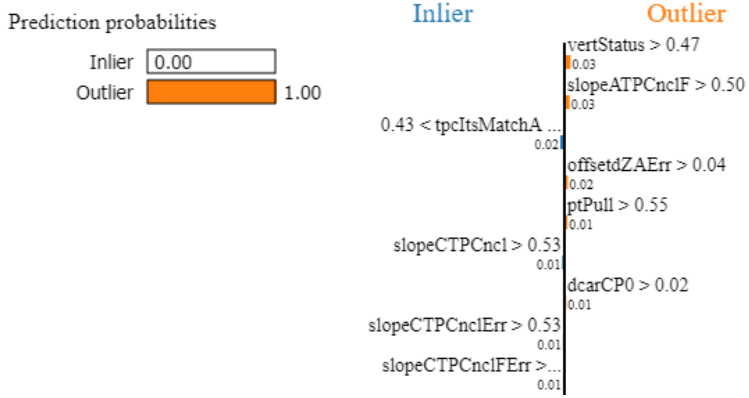


Feature	Value
errorMultPos	0.38
errorMultNeg	0.41
dcaz_negC_2_Err	0.79
dcar_posC_1_Err	0.85
offsetdZAErr	0.36
slopedZAErr	0.29
qOverPtC	1.00
dcar_posC_0_Err	0.82
dcaz_posC_0_Err	0.76
slopeATPCnclFErr	0.48
dcar_posA_0_Err	0.79
dcaz_posC_1_Err	0.81
dcar_posC_2_Err	0.80

LIME with Autoencoder



Feature	Value
slopeCTPCnclErr	0.43
slopeCTPCncl	0.43
slopeCTPCnclF	0.51
slopeCTPCnclFErr	0.51
tpcItsMatchA	0.53
slopeATPCnclF	0.57
vertStatus	0.43
meanMultPos	0.17
errorMultPos	0.03
meanTPCnclF	0.65
ptPull	0.49
medianHVandPTGainCorrOROC	0.73
offsetdZCchi2Pos	0.10



Feature	Value
vertStatus	0.67
slopeATPCnclF	0.63
tpcItsMatchA	0.53
offsetdZAErr	0.36
ptPull	0.58
slopeCTPCncl	0.53
dcarCP0	0.71
slopeCTPCnclErr	0.53
slopeCTPCnclFErr	0.72
meanMult	0.19
dcaz_posC_chi2	0.07
meanMultPos	0.19
meanPTRelativeA	0.70

Development of Tools for Data Quality Control for ALICE Experiment at CERN, Using Machine Learning Methods

Parameter	Random Forest	DBSCAN	Isolation Forest	Autoencoder
Accuracy on former method labels	Very high	High	High	High
Can work unsupervised	No	Yes	Yes	Yes
Generalization abilities on not labeled data	-	Low	Medium	High
Amount of hyperparams. and tuning difficulty	Low	High	Low	Very high
Training Time	Short	Short	Short	Long
Can find subgroups in data	No	Yes	No	No
Can outperform expert user?	No	Possibly	Possibly	Yes
Can be used with LIME?	Yes	Requires intermediate step	Requires intermediate step	Requires intermediate step but can also leverage native proper.

Bartłomiej Cerek