



LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

Big
ata
HEP

MACHINE LEARNING IN ANALYTICAL CHEMISTRY

Diogo Barros, Pier Parpot

11th janurary 2019

Coimbra, PT

Work Context

Study object:

- PCBs (Printed Circuit Boards) supplied by Bosch Car Multimedia.
- Train a model to classify PCBs as (not) contaminated fed with data obtained from chemical analysis.



Experimental Procedure

Extraction

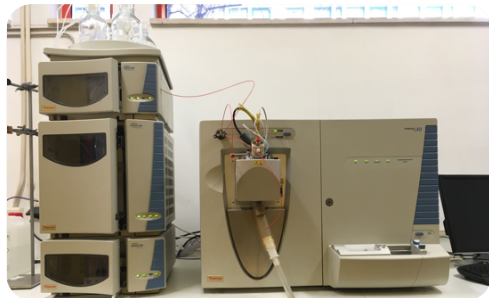
Solid-Liquid
Extraction
(SLE)

IPC-TM-650 2.3.27.1



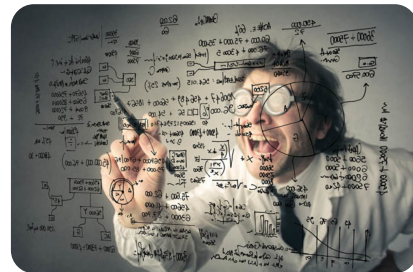
Chemical Analysis

High Performance
Liquid Chromatography
Mass Spectrometry
(HPLC-MS)

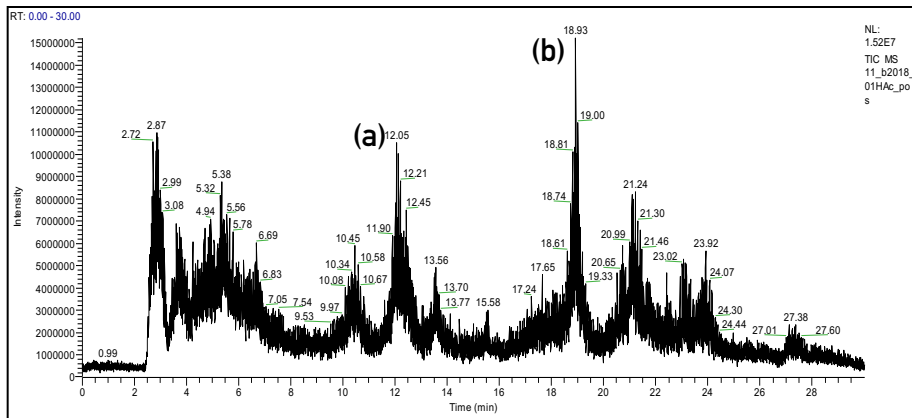


Data treatment

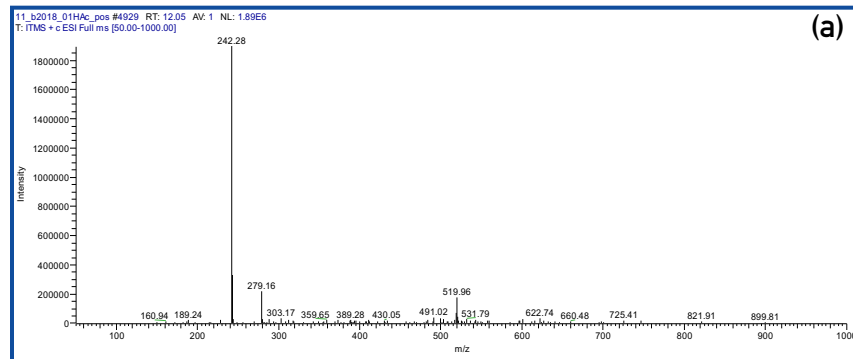
Data mining
(Python)



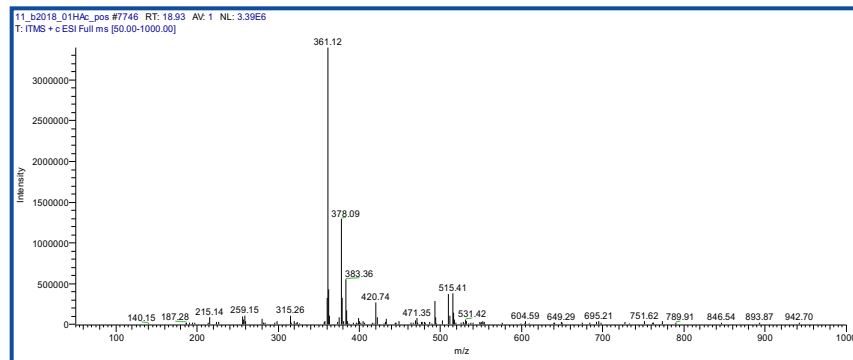
Data Wrangling



~12000 datapoints



950 datapoints

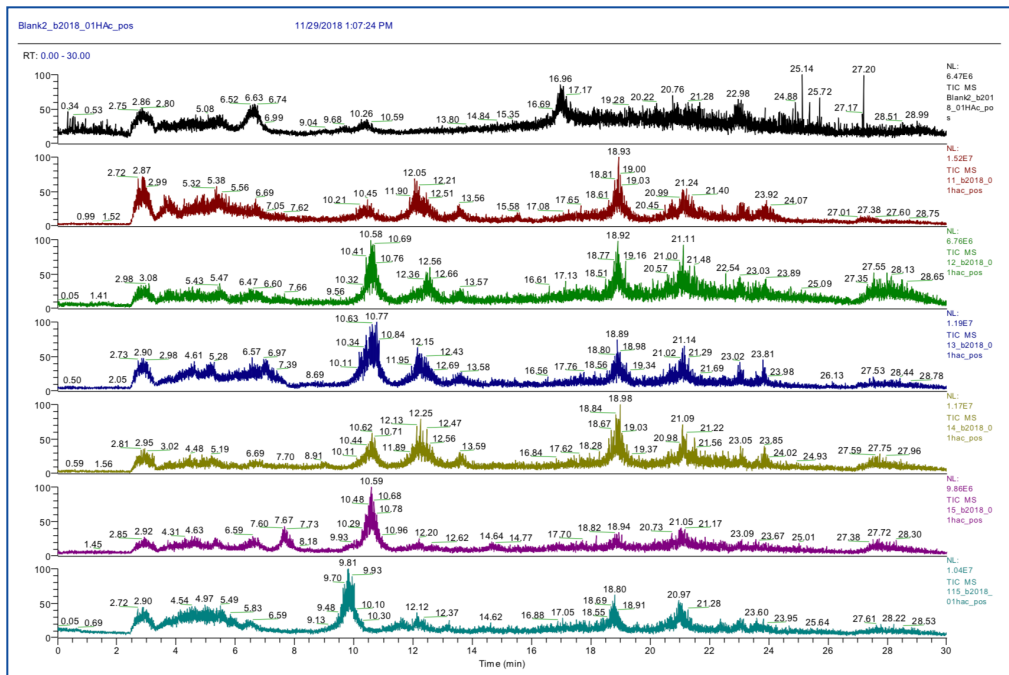


Data Wrangling

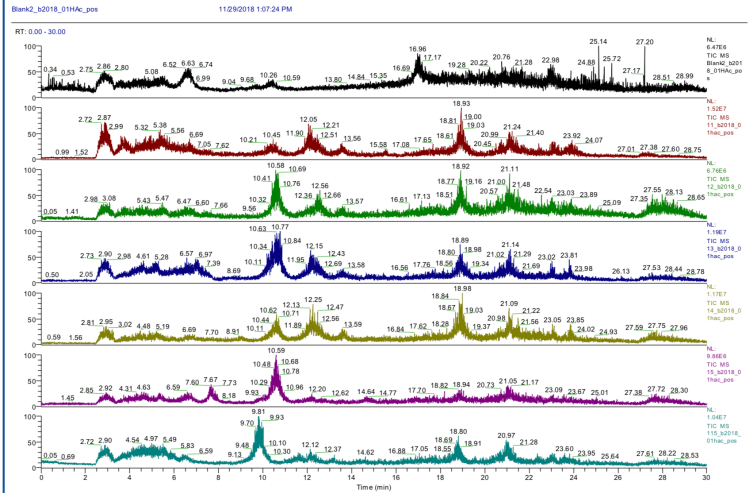
	50	51	52	53	54	55	56	57	58	59	...	991	992	993	994	995	996	997	998
0	22194	54725	21432	7910	408	90909	47055	64532	23176	50642	...	19034	19126	1721	16164	64033	46383	73387	4137
1	96227	32921	90813	9374	67164	55385	18811	74052	10945	39086	...	89898	79698	60540	42823	37038	70972	51913	22631
2	43253	98227	20031	76528	40598	20732	7856	34572	80996	20931	...	94306	34163	44239	61439	36294	60300	16496	66500
3	66278	83296	45076	14495	41301	52857	47902	38288	19965	60547	...	99341	91337	81978	49671	99973	56892	45883	36167
4	18201	34492	82416	47035	3737	11428	41301	43274	74740	33872	...	25198	14919	68088	55157	30244	59379	77629	85870
5	56285	65556	91780	32155	4120	45320	92580	10656	7657	14457	...	6292	54773	5171	89798	64312	28901	71259	17418
6	36842	48068	23478	58861	81182	18809	26765	38691	53412	7367	...	65128	65664	89107	97211	40890	42233	84013	90044
7	81314	71813	18504	5282	34578	14936	69730	47622	25250	93291	...	94784	68013	74516	51159	65067	4282	75492	21842
8	55415	5290	72997	95797	16774	27929	90370	87378	29948	46482	...	5200	76825	48357	56534	19549	98050	92184	12255
9	91253	30867	24060	10362	84048	15746	15182	5511	90591	9119	...	26231	44207	2034	26775	51389	44999	52377	1337
10	12946	54637	24689	76598	61284	42400	52863	67203	62052	78796	...	89508	50691	10719	57400	2893	93066	15828	33619
11	21556	18138	95818	19968	83310	48075	81839	42176	42169	1356	...	48419	67448	24095	98232	82407	25652	97634	89049
12	32996	86762	89656	97282	81609	20881	34805	4964	92435	18797	...	39887	32836	60086	85993	54903	47519	99628	72386
13	67260	26963	1521	35075	36511	9237	42951	33290	97186	17087	...	745	50515	67523	1609	82142	9459	70940	31276
14	49484	98851	38264	17939	61783	33132	84859	15029	99420	21867	...	8366	21581	62778	7427	35862	31100	89175	47775

- $x = \text{masses (50 - 1000)Da}$
- $y = \text{time (12k datapoints)}$
- **Shape (950, 12000)**

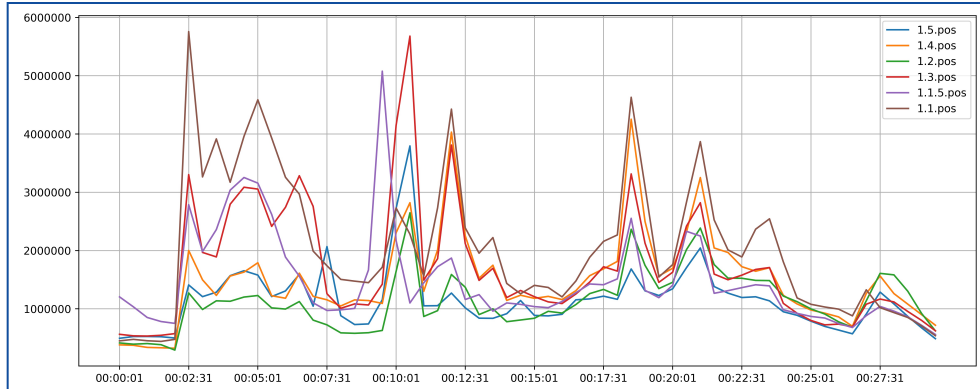
Data mining “brute force” approach



- Each chromatogram represents a sample
- Visible pattern among samples
- Seems clever to use time as predictive feature



~12000 datapoints

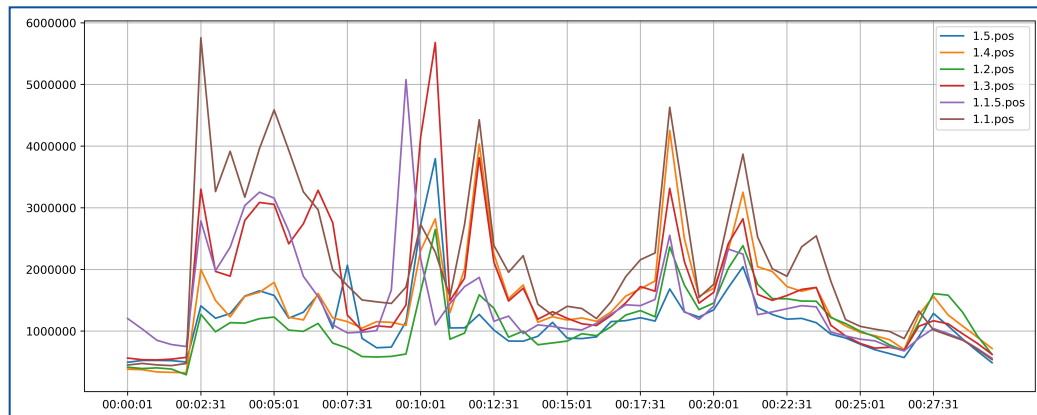


60 datapoints

[illegible]

Data mining

Feature engineering approach



	00:00:01	00:00:31	00:01:01	00:01:31	00:02:01	00:02:31	00:03:01	00:03:31	00:04:01
1.5.pos	4.948382e+05	5.249579e+05	525739.883333	521650.775397	499838.630159	1.410234e+06	1.209399e+06	1.285888e+06	1.569353e+06
1.4.pos	3.820739e+05	3.748238e+05	338593.482540	332126.689683	324048.813492	2.001369e+06	1.498057e+06	1.230812e+06	1.562850e+06
1.2.pos	4.148494e+05	3.938719e+05	404375.784127	384809.158730	292515.057143	1.272567e+06	9.890456e+05	1.137694e+06	1.129999e+06
1.3.pos	5.633455e+05	5.368122e+05	533404.807143	546116.104762	574996.476984	3.303113e+06	1.968239e+06	1.891453e+06	2.795662e+06
1.1.5.pos	1.205078e+06	1.034886e+06	852045.073016	782434.719841	750664.167460	2.787587e+06	1.982766e+06	2.360404e+06	3.037262e+06

5 rows × 60 columns

Create new features from the original dataset:

- n biggest peaks;
- Distance between peaks;
- Intensity ratio of n/n+1 biggest peaks;
- Sum of the intensities of n peaks;
- (...)

Data Augmentation

Extraction

**Solid-Liquid
Extraction
(SLE)**

IPC-TM-650 2.3.27.1



Chemical Analysis

**High
Performance
Liquid
Chromatography
Mass
Spectrometry
(HPLC-MS)**



Data treatment

**Data mining
(Python)**

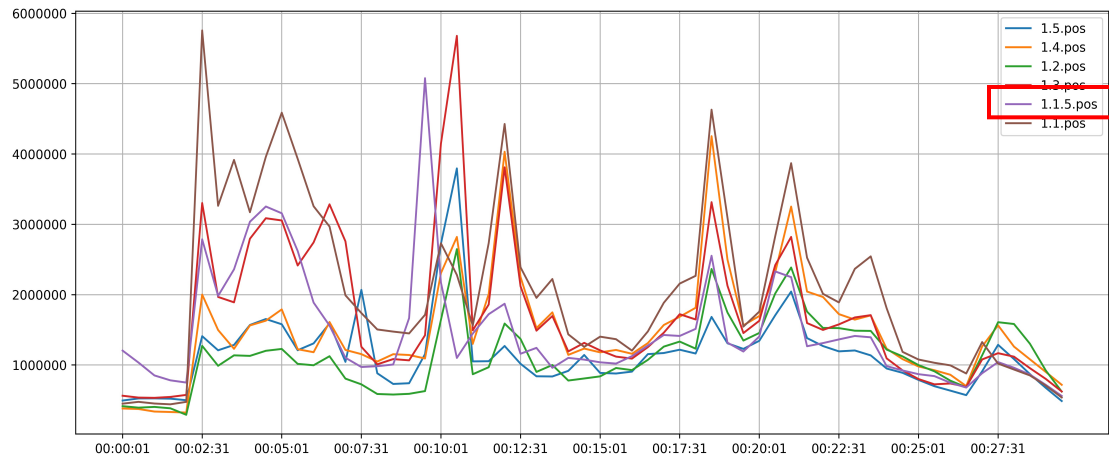
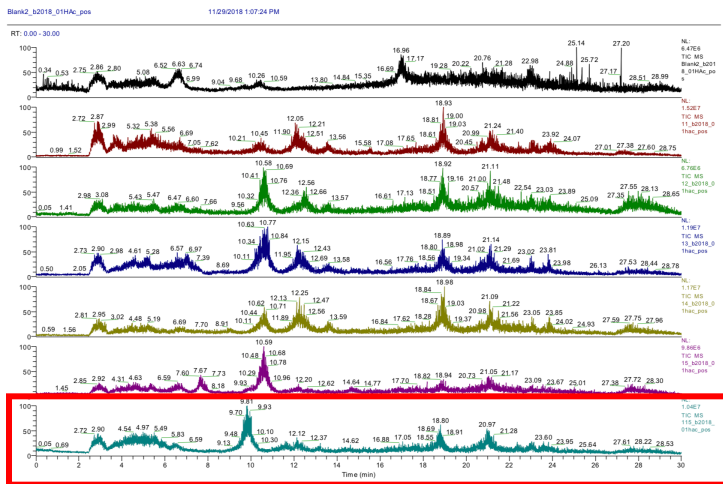


**1 mL
PCB A
(real)**

**1 mL
PCB B
(real)**

**PCB AB
(1:1)
(virtual)**

Data Augmentation

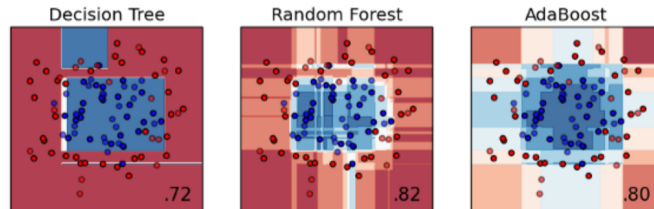


- Virtual sample exhibits chemical properties from both 'parents'
- Good potential to be used as a data augmentation technique

Machine Learning Models

ML implementations from SKLearn:

- Random Forests
- Boosted Decision Trees (XGBoost, AdaBoost)





LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

Thanks!

Any questions?

diogobarrosgoncalves@gmail.com