Computational Notebooks for Cheminformatics

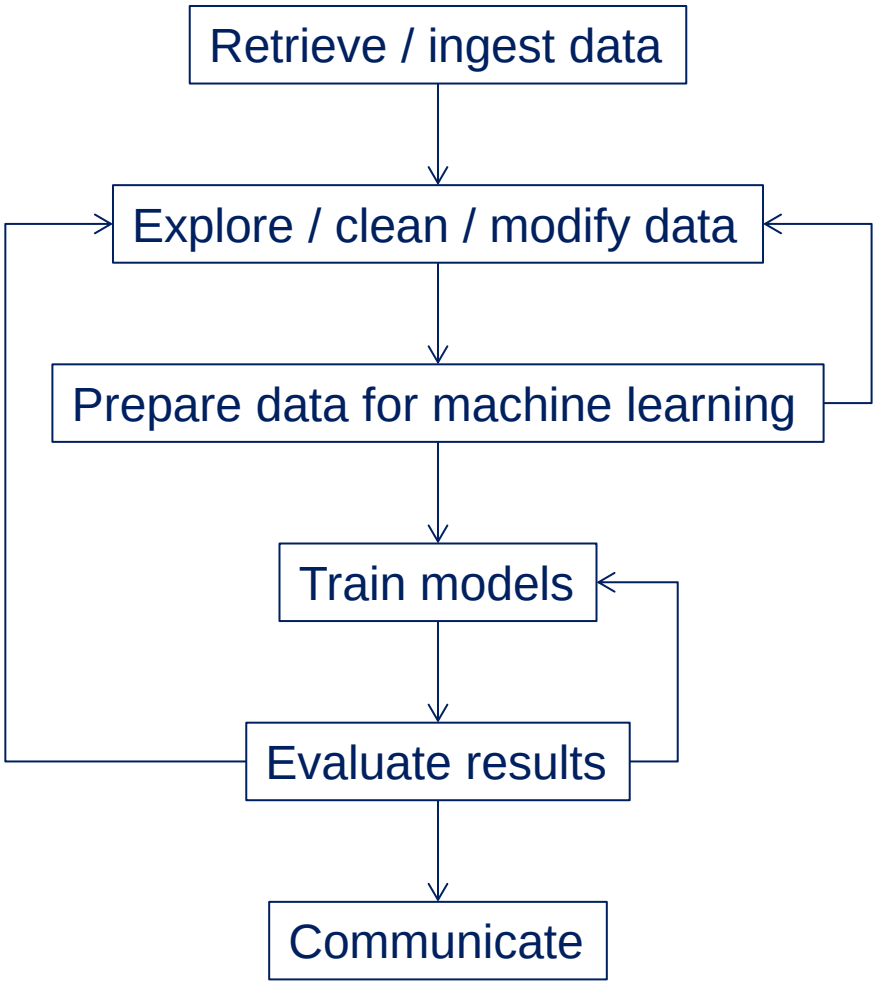
ACS Fall 2019 CHED 285

Paul J Kowalczyk
Senior Data Scientist

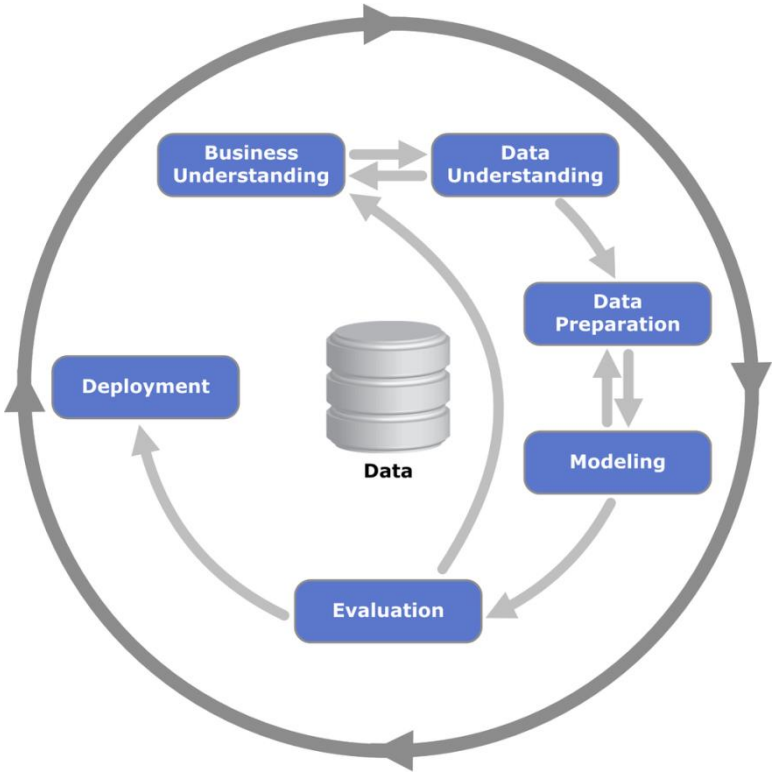
paul.kowalczyk@solvay.com
www.linkedin.com/in/PaulJKowalczyk

R&I Centre Bristol

Machine Learning Workflow

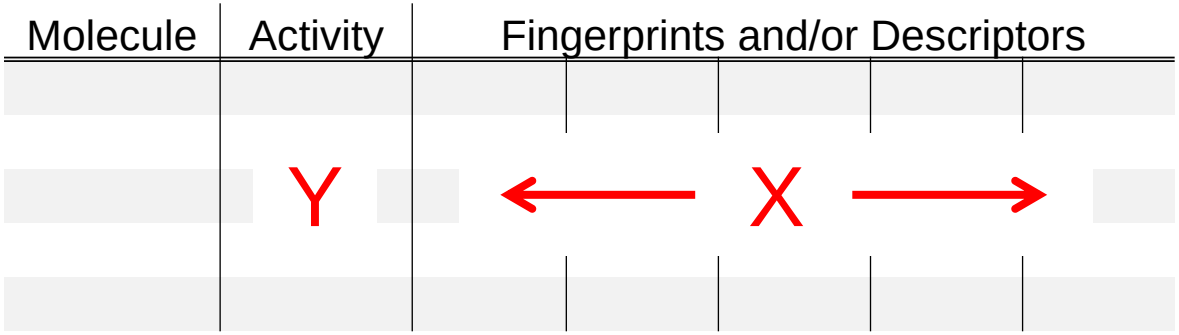


CRoss-Industry Standard Process for Data Mining

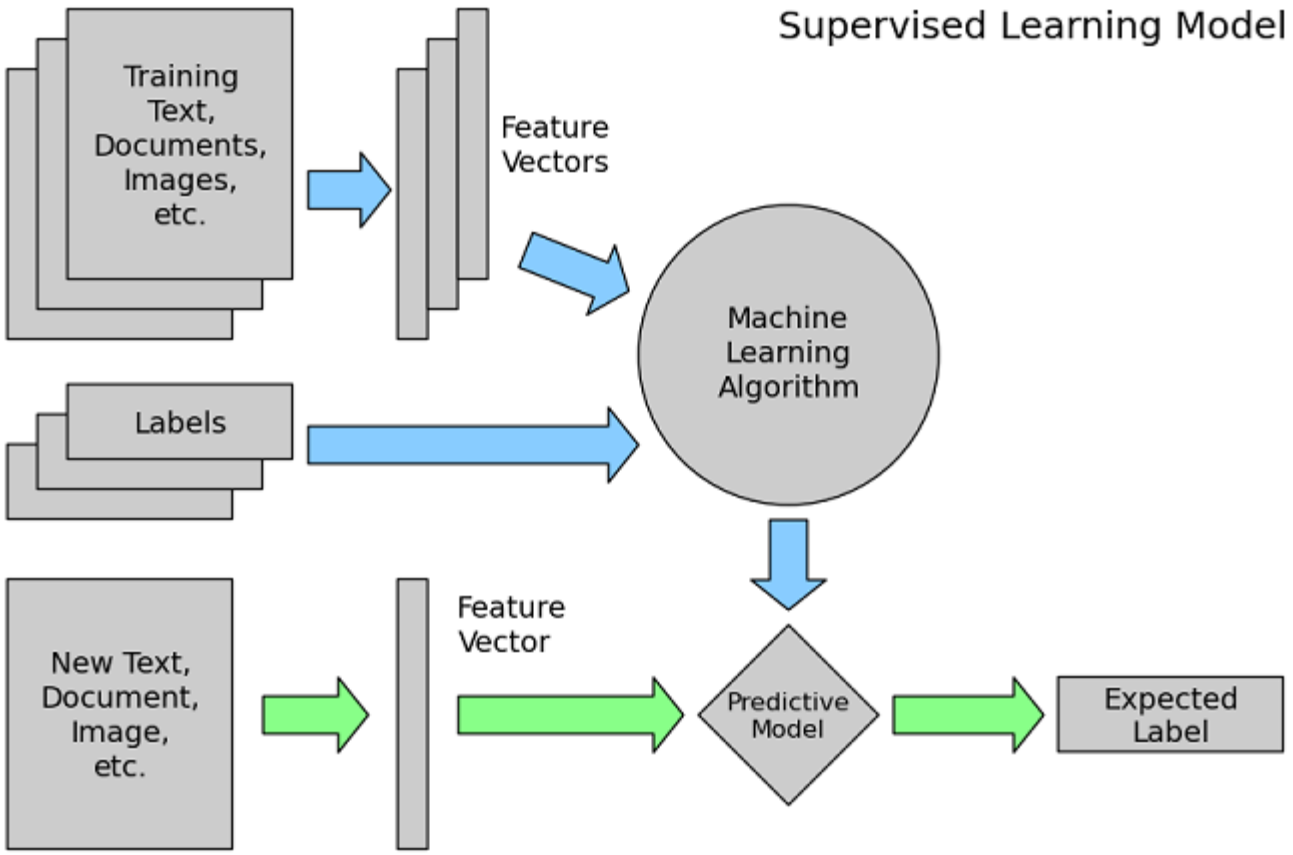


Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts.

The dataframe

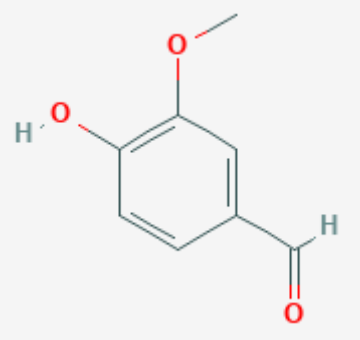


Supervised Learning Model



Representing molecules: vanillin

Vanillin



Canonocal SMILES:

COC1=C(C=CC(=C1)C=O)O

InChI=1S/C8H8O3/c1-11-8-4-6(5-9)2-3-7(8)10/h2-5,10H,1H3

InChI Key:

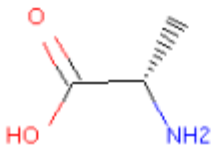
MWOOGOJBHIARFG-UHFFFAOYSA-N

Representing molecules: alanine.mol

```
In [37]: alanine = Chem.MolFromInchi('InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1')

In [39]: Draw.MolToImage(alanine)

Out[39]:
```

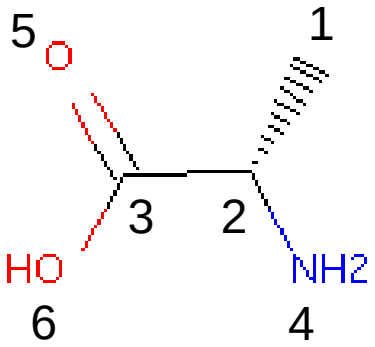



```
In [40]: Chem.MolToMolFile(m1, 'alanine.mol')
```

```
     RDKit          2D

  6  5  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2990    0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5981   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2990    2.2500    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.5981   -1.5000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.8971    0.7500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  6
  2  3  1  0
  2  4  1  0
  3  5  2  0
  3  6  1  0
M  END
```

Representing molecules: alanine.mol




```
     RDKit          2D

  6  5  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2990    0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.5981   -0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2990    2.2500    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.5981   -1.5000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.8971    0.7500    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  6
  2  3  1  0
  2  4  1  0
  3  5  2  0
  3  6  1  0
M  END
```

Prepare data for machine learning: Descriptors (features)

List of Available Descriptors

Descriptor/Descriptor Family	Notes
Gasteiger/Marsili Partial Charges	*Tetrahedron* **36**:3219–28 (1980)
BalabanJ	*Chem. Phys. Lett.* **89**:399–404 (1982)
BertzCT	*J. Am. Chem. Soc.* **103**:3599–601 (1981)
Ipc	*J. Chem. Phys.* **67**:4517–33 (1977)
HallKierAlpha	*Rev. Comput. Chem.* **2**:367–422 (1991)
Kappa1 – Kappa3	*Rev. Comput. Chem.* **2**:367–422 (1991)
Chi0, Chi1	*Rev. Comput. Chem.* **2**:367–422 (1991)
Chi0n – Chi4n	*Rev. Comput. Chem.* **2**:367–422 (1991)
Chi0v – Chi4v	*Rev. Comput. Chem.* **2**:367–422 (1991)
MolLogP	Wildman and Crippen *JCICS* **39**:868–73 (1999)
MolMR	Wildman and Crippen *JCICS* **39**:868–73 (1999)
MolWt	
ExactMolWt	
HeavyAtomCount	
HeavyAtomMolWt	
NHOHCount	
NOCount	
NumHAcceptors	
NumHDonors	
NumHeteroatoms	
NumRotatableBonds	
NumValenceElectrons	
NumAmideBonds	
Num{Aromatic,Saturated,Aliphatic}Rings	
Num{Aromatic,Saturated,Aliphatic} {Hetero,Carbo}cycles	
RingCount	
FractionCSP3	

NumSpiroAtoms	Number of spiro atoms (atoms shared between rings that share exactly one atom)
NumBridgeheadAtoms	Number of bridgehead atoms (atoms shared between rings that share at least two bonds)
TPSA	*J. Med. Chem.* **43**:3714–7, (2000) See the section in the RDKit book describing differences to the original publication.
LabuteASA	*J. Mol. Graph. Mod.* **18**:464–77 (2000)
PEOE_VSA1 – PEOE_VSA14	MOE-type descriptors using partial charges and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm
SMR_VSA1 – SMR_VSA10	MOE-type descriptors using MR contributions and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm
SlogP_VSA1 – SlogP_VSA12	MOE-type descriptors using LogP contributions and surface area contributions http://www.chemcomp.com/journal/vsadesc.htm
EState_VSA1 – EState_VSA11	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper)
VSA_EState1 – VSA_EState10	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper)
MQNs	Nguyen et al. *ChemMedChem* **4**:1803–5 (2009)
Topliss fragments	implemented using a set of SMARTS definitions in $(RDBASE)/Data/FragmentDescriptors.csv
Autocorr2D	New in 2017.09 release. Todeschini and Consoni "Descriptors from Molecular Geometry" Handbook of Chemoinformatics http://dx.doi.org/10.1002/9783527618279.ch37

Prepare data for machine learning: Fingerprints

List of Available Fingerprints

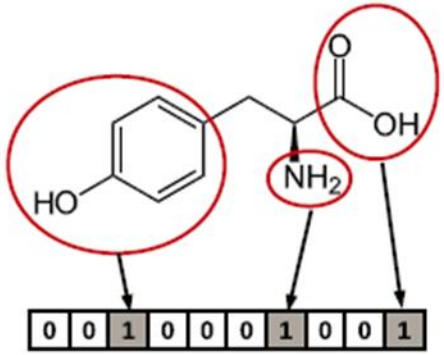
Fingerprint Type	Notes
RDKit	a Daylight-like fingerprint based on hashing molecular subgraphs
Atom Pairs	*JCICS* **25**:64-73 (1985)
Topological Torsions	*JCICS* **27**:82-5 (1987)
MACCS keys	Using the 166 public keys implemented as SMARTS
Morgan/Circular	Fingerprints based on the Morgan algorithm, similar to the ECFP/FCFP fingerprints *JCIM* **50**:742-54 (2010).
2D Pharmacophore	Uses topological distances between pharmacophoric points.
Pattern	a topological fingerprint optimized for substructure screening
Extended Reduced Graphs	Derived from the ErG fingerprint published by Stiefl et al. in *JCIM* **46**:208-20 (2006). NOTE: these functions return an array of floats, not the usual fingerprint types

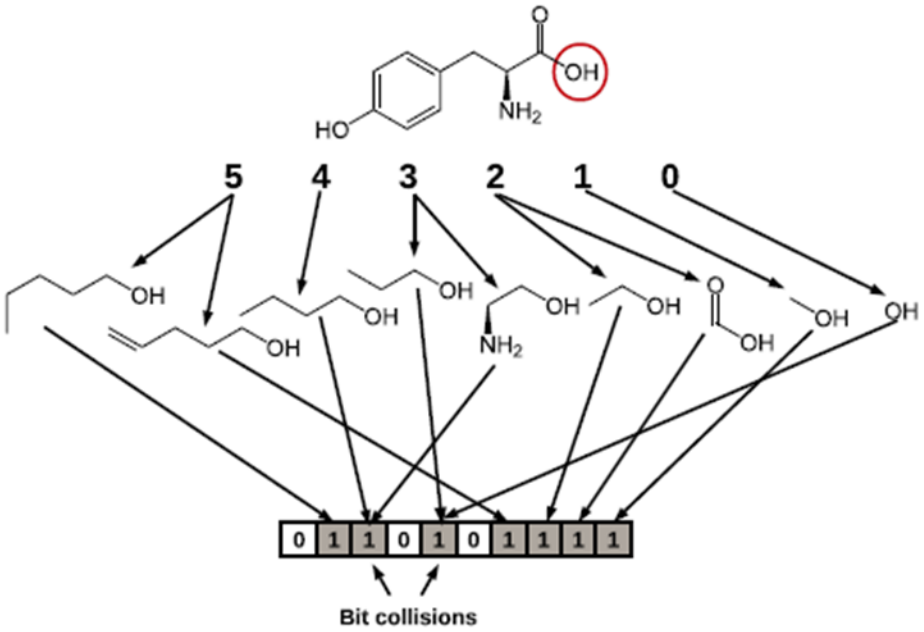
Feature Definitions Used in the Morgan Fingerprints

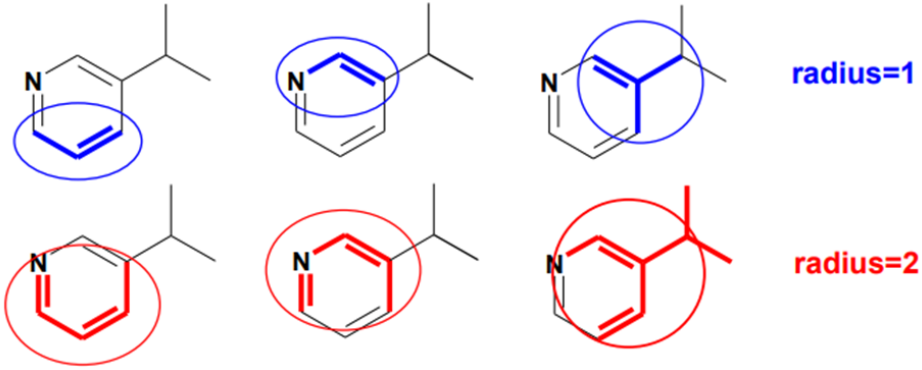
These are adapted from the definitions in Gobbi, A. & Poppinger, D. "Genetic optimization of combinatorial libraries." *Biotechnology and Bioengineering* **61**, 47-54 (1998).

Feature	SMARTS
Donor	`[$([N;!H0;v3,v4&+1]),$([O,S;H1;+0]),n&H1&+0]`
Acceptor	`[$([O,S;H1;v2;!$(*-*=[O,N,P,S])]),$([O,S;H0;v2]),$([O,S;-]),$([N;v3;!$(N-*=[O,N,P,S])]),n&H0&+0,$([o,s;+0;!$([o,s]:n);!$([o,s]:c:n))]`
Aromatic	`[a]`
Halogen	`[F,Cl,Br,I]`
Basic	`[#7;+,$([N;H2&+0][$([C,a]);!$([C,a](=O))]),$([N;H1&+0]([$([C,a]);!$([C,a](=O))])[$([C,a]);!$([C,a](=O))]),$([N;H0&+0]([C;!$(C(=O))])([C;!$(C(=O))])[C;!$(C(=O))])]`
Acidic	`[$([C,S](=[O,S,P])-[O;H1,-1])]`

Prepare data for machine learning: Fingerprints







Evaluate results: Cohen's kappa

kappa takes into account the possibility of the agreement occurring by chance

	Reference	
Predicted	Event	No Event
Event	A	B
No Event	C	D

A: true positive
B: false positive (Type I error)
C: false negative (Type II error)
D: true negative

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

$$\Pr(a) = \frac{A+D}{A+B+C+D}$$

$$P_{YES} = \frac{A+B}{A+B+C+D} \times \frac{A+C}{A+B+C+D} \qquad P_{NO} = \frac{C+D}{A+B+C+D} \times \frac{B+D}{A+B+C+D}$$

$$\Pr(e) = P_{YES} + P_{NO}$$

Evaluate results: Cohen's kappa

Kappa	Agreement
< 0	Less than chance agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Almost perfect agreement

Retrieve / ingest data

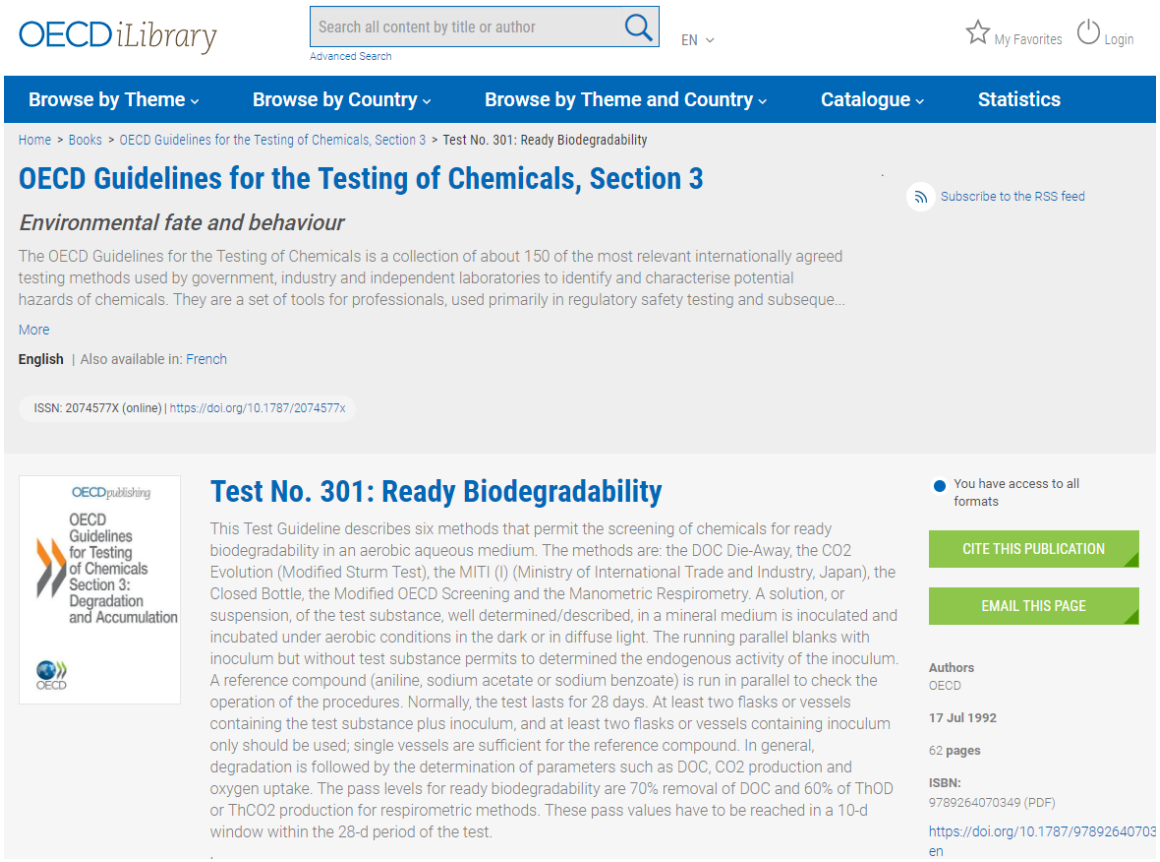
Table 1 Endpoint datasets in the PHYSPROP database

Property abbreviation	Property
AOH	Atmospheric hydroxylation rate
BCF	Bioconcentration factor
BioHL	Biodegradability half-life
BP	Boiling point
HL	Henry's Law constant
KM	Fish biotransformation half-life
KOA	Octanol–air partition coefficient
KOC	Soil adsorption coefficient
logP	Octanol–water partition coefficient
MP	Melting point
RB	Readily biodegradable
VP	Vapor pressure
WS	Water solubility

Mansouri, K., Grulke, C. M., Judson, R. S., & Williams, A. J. (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. *Journal of cheminformatics*, *10*(1), 10.

OPEn structure-activity/property Relationship App

Ready Biodegradability

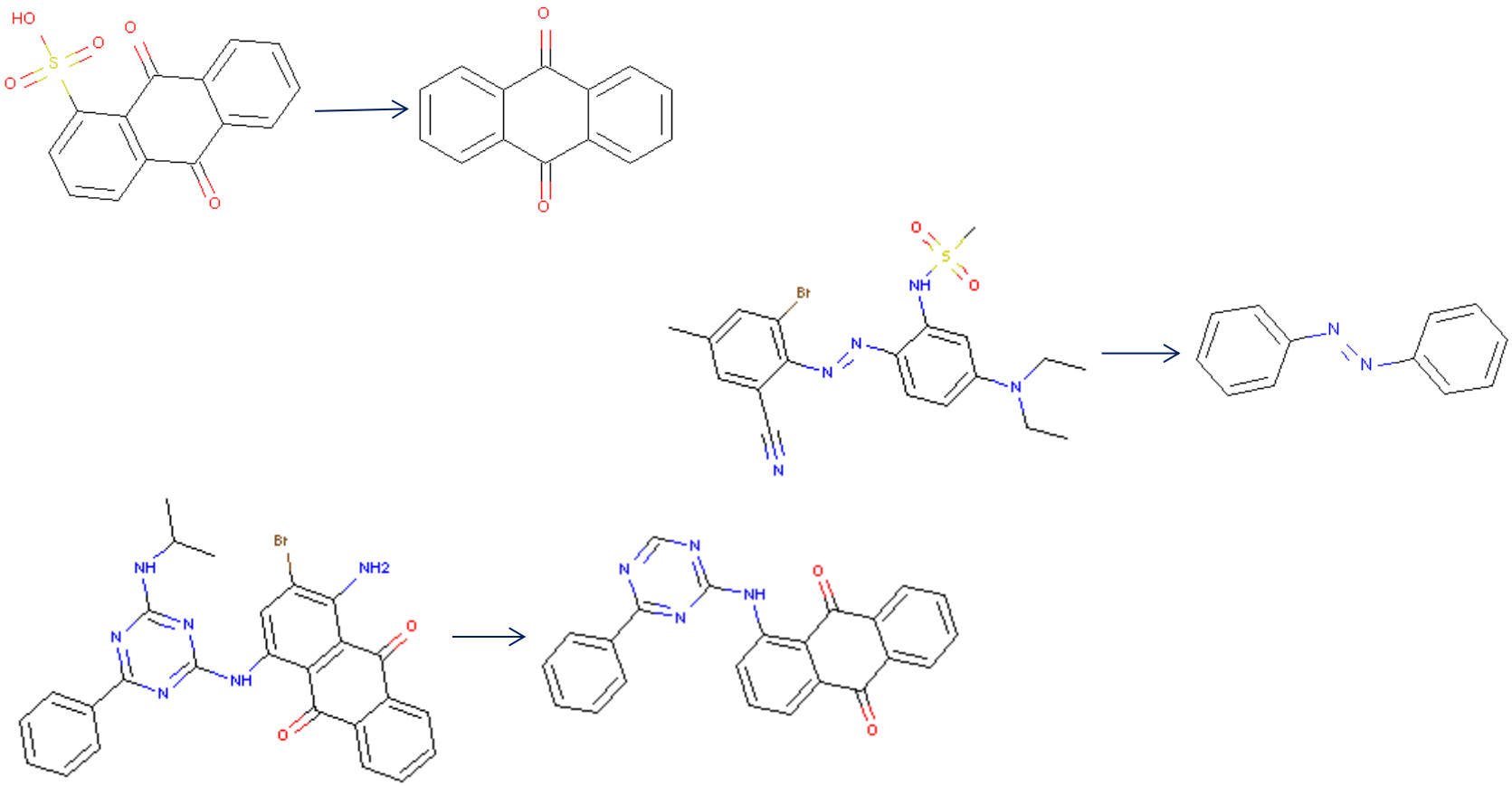


Ready Biodegradable (RB): 681
Not Ready Biodegradable (NRB): 1304

Explore / clean / modify data: structure curation

- Ingest the 3 biodegradability datasets
    - Cheng (JChemInfModel_52_655)
    - Mansouri (JCIM_53_867)
    - OPERA (OPERA)
- Sanitize molecules
- Identify replicates in the datasets
- Compare / contrast the datasets, *e.g.,* compare molecular weight distributions, TPSA, logP, …

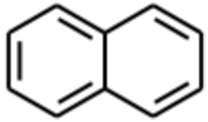
Explore / clean / modify data: Murcko frameworks



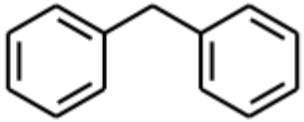
Bemis, Guy W., and Mark A. Murcko. "The properties of known drugs. 1. Molecular frameworks." *Journal of medicinal chemistry* 39.15 (1996): 2887-2893.

Explore / clean / modify data: Murcko frameworks

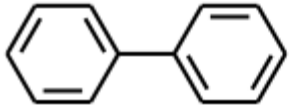
512



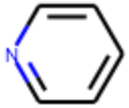
43



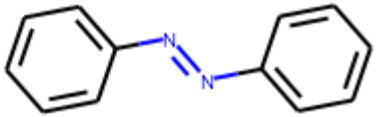
36



34

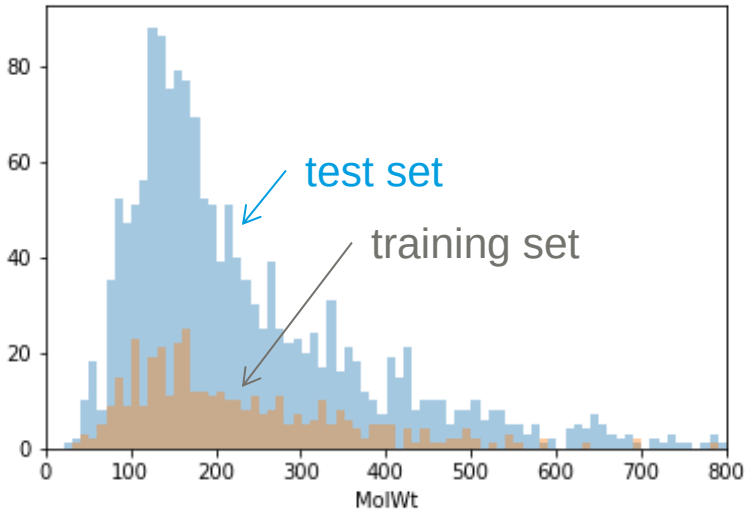


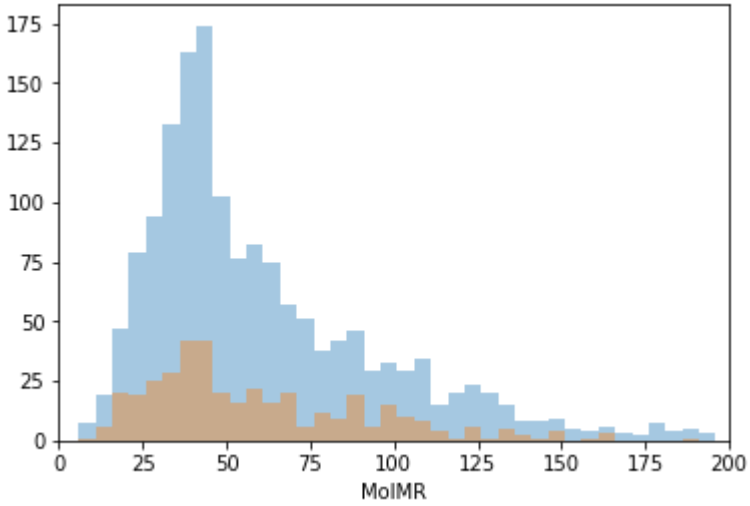
34

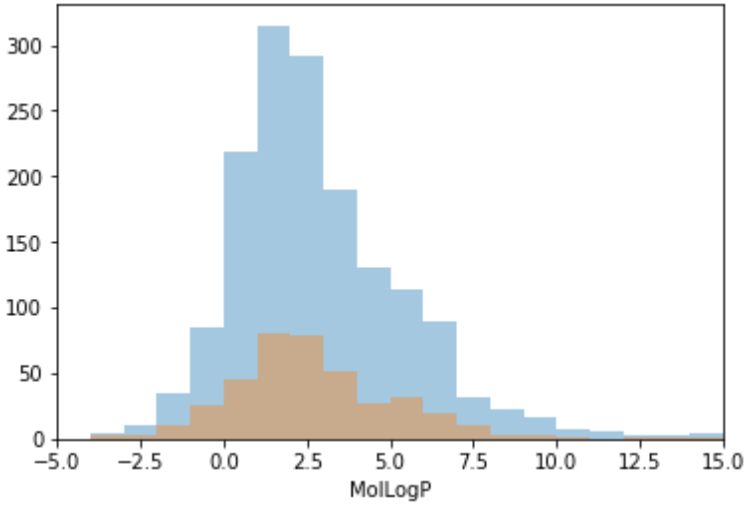


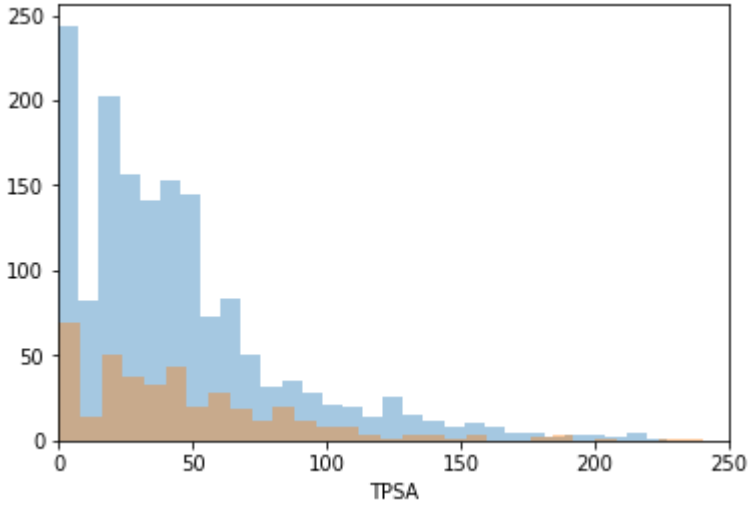
28

Explore / clean / modify data









Model: recursive partitioning / random forests

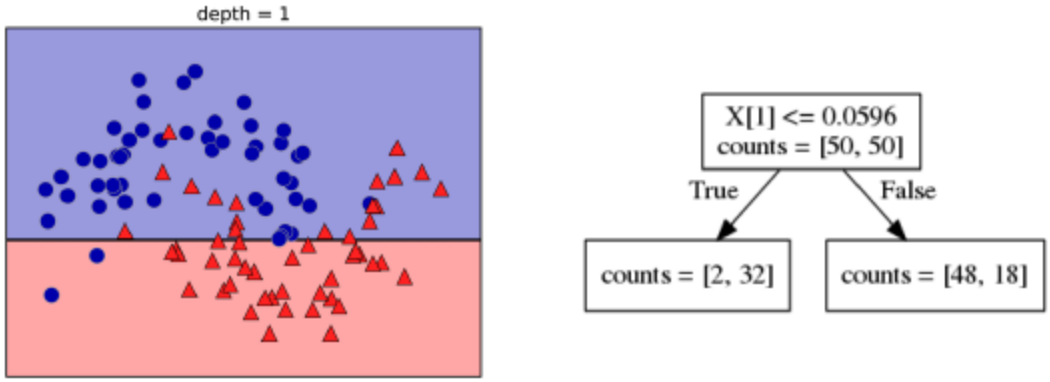


Figure 2-24. Decision boundary of tree with depth 1 (left) and corresponding tree (right)

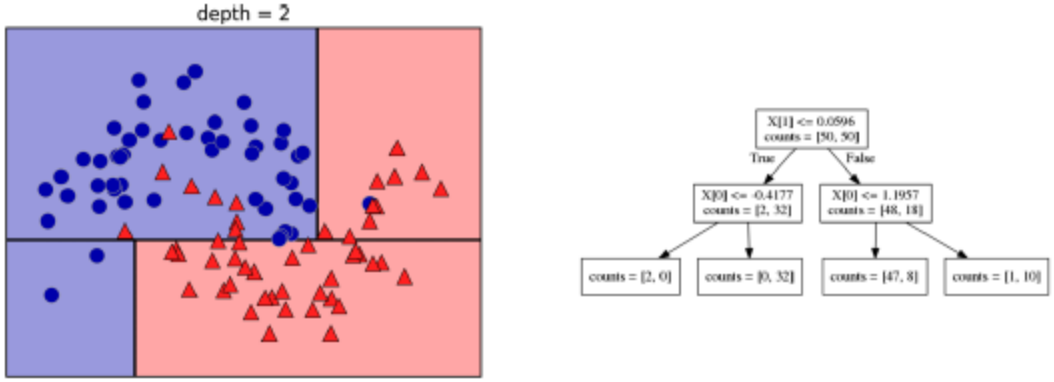
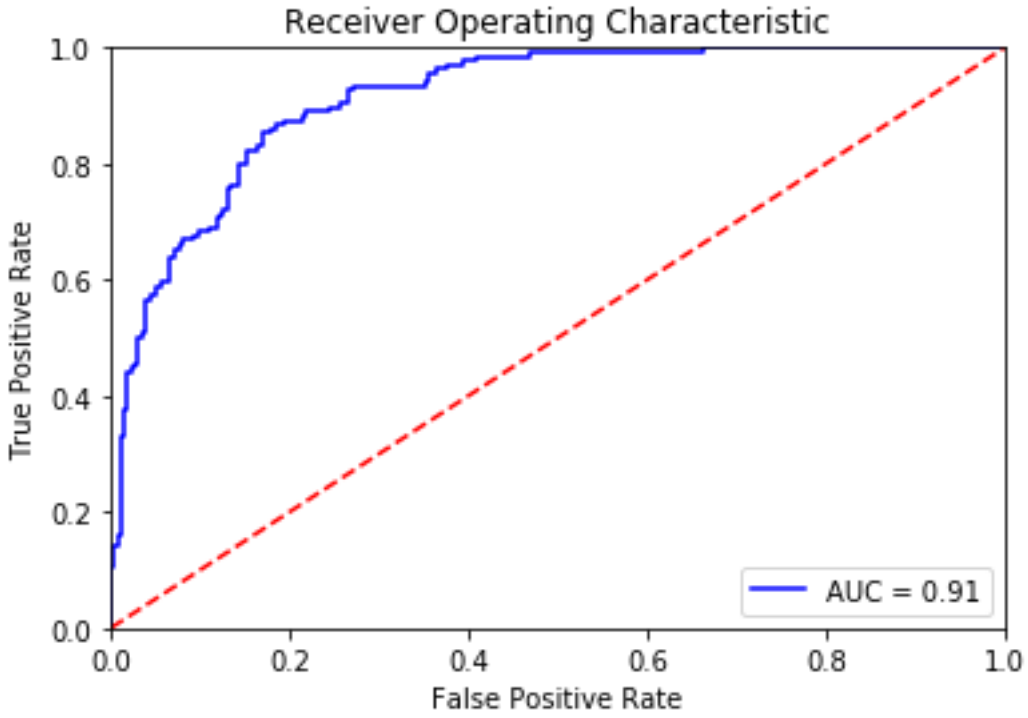


Figure 2-25. Decision boundary of tree with depth 2 (left) and corresponding decision tree (right)

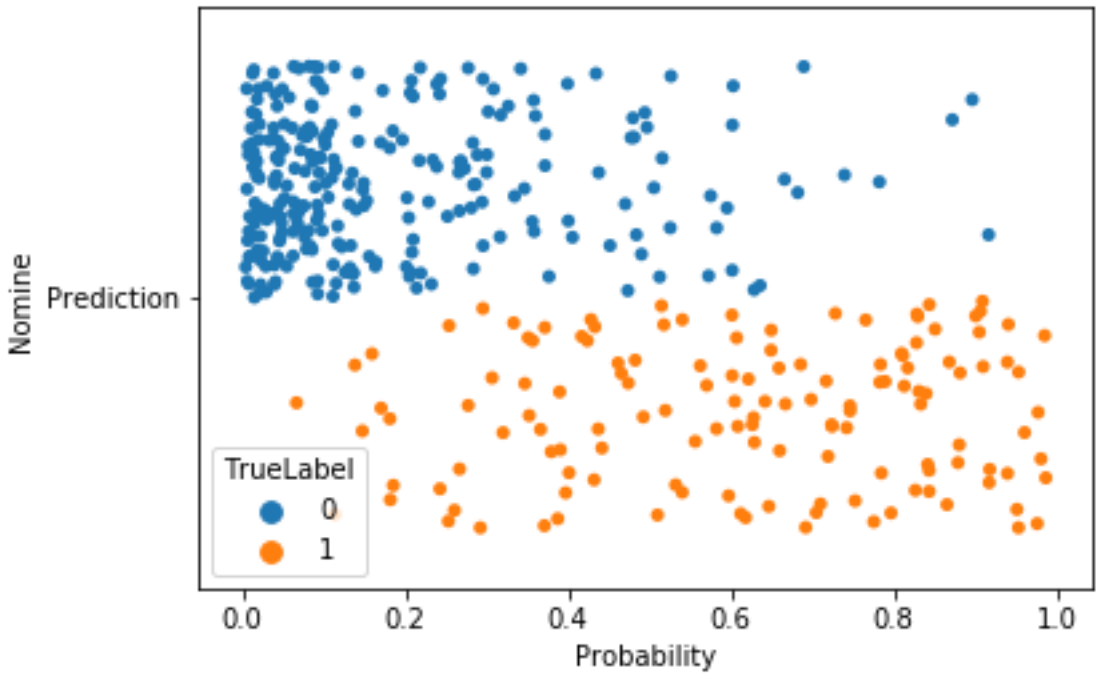
Model Performance
ROC & AUC



Model Performance: Confusion Matrix

kappa = 0.61

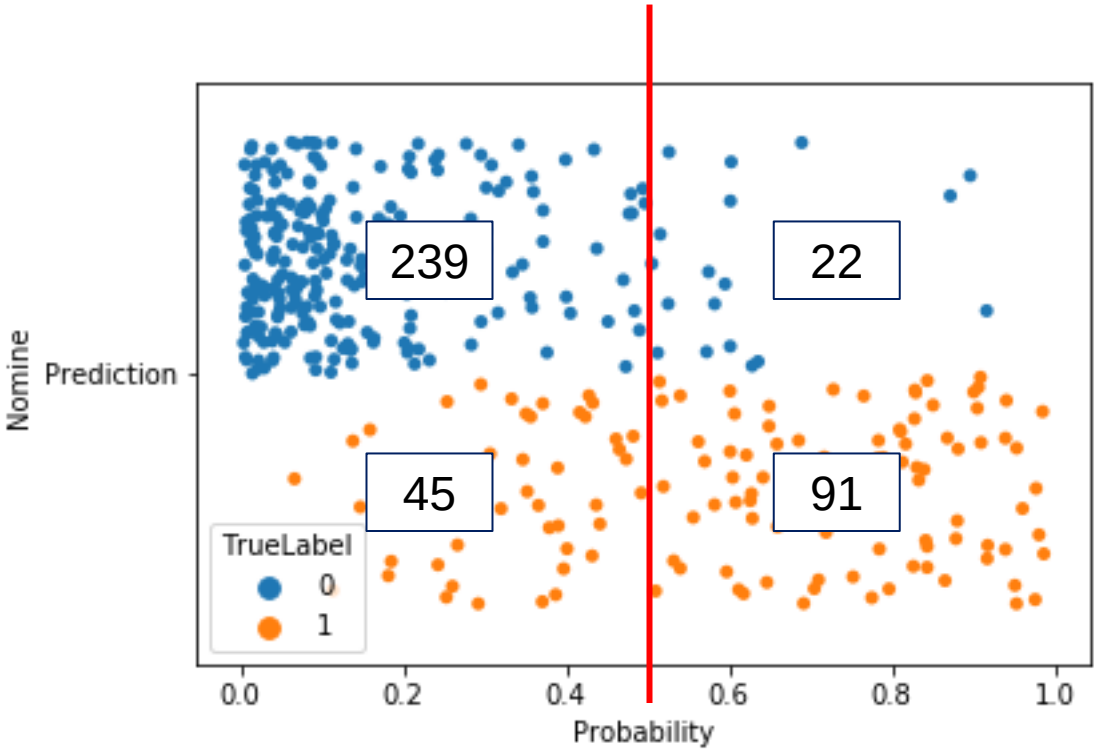
```
[[239  22]
 [ 45  91]]
```



Model Performance: Confusion Matrix

kappa = 0.61

[[239 22]
[45 91]]



notebook

Merci!

Thank You!

Dziękuję

paul.kowalczyk@solvay.com
www.linkedin.com/in/PaulJKowalczyk

www.solvay.com

SOLVAY
asking more from chemistry®