(Machine) Learning on Open Datasets

Giulia Santarsieri, Pavel Soriano-Morales

May 5, 2021

csv,conf,v6

AI Lab – Etalab – Direction Intérministerielle du numérique

Etalab’s mission to uncover the potential of Open Data

etalab gouv.fr

data.gouv.fr

Promote Open Data

Lab IA

Exploit Open Data with Data Science and AI

Support data driven public policy

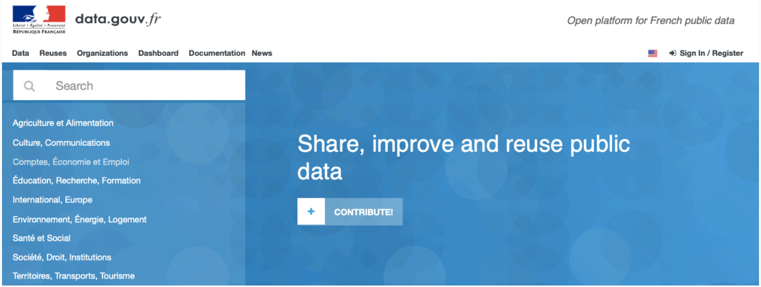
Table of contents

1. The lack of Open Data in Machine Learning

2. Our methodology for a ML data repository with Open Data

3. DGML: Data Gouv for Machine Learning

The lack of Open Data in Machine Learning


Liberté • Égalité • Fraternité
République Française
data.gouv.fr
Data
Reuses
Organizations
Dashboard
Documentation
News
Sign In / Register
Search
Agriculture et Alimentation
Culture, Communications
Comptes, Économie et Emploi
Éducation, Recherche, Formation
International, Europe
Environnement, Énergie, Logement
Santé et Social
Société, Droit, Institutions
Territoires, Transports, Tourisme
+ CONTRIBUTE!

Open platform for French public data

Share, improve and reuse public data

Linkage of Hospital Records and Death Certificates by a Search Engine and Machine Learning

Published on March 1, 2021 by **Equipe de recherche en informatique appliquée à la santé**

Application pour relier les données hospitalières aux certificats de décès développée au CHU de Bordeaux.

Deep learning pour la prédiction de la densité

Publié le 3 juin 2020 par **Etienne Kintzler**

Modèle de deep learning appliqué à la prédiction de la densité de population à partir des carreaux INSEE et des images satellites (base BD ORTHO de l'IGN).

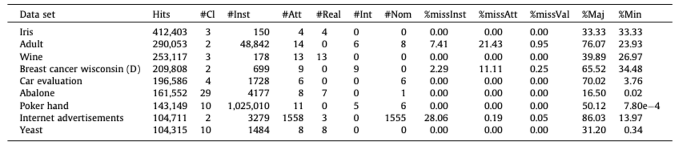
Machine learning pour prédire la gravité des accidents

Publié le 19 septembre 2020 par **Ilyes Talbi**

Dans cet article, j'utilise une base de données qui recense des accidents de la circulation pour créer un modèle de machine learning. Ce projet sera l'occasion d'introduire Random Forest et XGBoost et de comparer leurs performances, à travers un tutoriel pas à pas.

Data for Machine Learning

A **small number** of well-known datasets is often used[1] in Machine Learning research and applications



Data set	Hits	#Cl	#Inst	#Att	#Real	#Int	#Nom	%missInst	%missAtt	%missVal	%Maj	%Min
Iris	412,403	3	150	4	4	0	0	0.00	0.00	0.00	33.33	33.33
Adult	290,053	2	48,842	14	0	6	8	7.41	21.43	0.95	76.07	23.93
Wine	253,117	3	178	13	13	0	0	0.00	0.00	0.00	39.89	26.97
Breast cancer wisconsin (D)	209,808	2	699	9	0	9	0	2.29	11.11	0.25	65.52	34.48
Car evaluation	196,586	4	1728	6	0	0	6	0.00	0.00	0.00	70.02	3.76
Abalone	161,552	29	4177	8	7	0	1	0.00	0.00	0.00	16.50	0.02
Poker hand	143,149	10	1,025,010	11	0	5	6	0.00	0.00	0.00	50.12	7.80e−4
Internet advertisements	104,711	2	3279	1558	3	0	1555	28.06	0.19	0.05	86.03	13.97
Yeast	104,315	10	1484	8	8	0	0	0.00	0.00	0.00	31.20	0.34

[1] Núria Macià et al. "Learner excellence biased by data set selection: A case for data characterisation and artificial data sets". In: *Pattern Recognition* 46.3 (2013), pp. 1054–1066.

Data for Machine Learning

These datasets do not always reflect the **challenges** of Open Data:

Code AGB	Food name	Environmental score
19580	Apricot, canned in light syrup, drained	2.46
NAN	Apricot, canned in light syrup, not drained	NAN
21508	Apricot, pitted, raw	2.5
21546	Caribbean-style fish fritters, fish acras	NAN
36780	NAN	2.46
25263	Yogurt, fermented milk or dairy specialty	3.61
NAN	Yogurt, fermented milk or dairy specialty	2.5
90768	Lamb, neck, raw	2.1

The advantages of using Open Data

Evaluate and challenge Machine Learning algorithms

Machine Learning for education and research

Machine Learning for business

Machine Learning to support public policy

Why is Open Data neglected in Machine Learning applications?

Lack of data quality

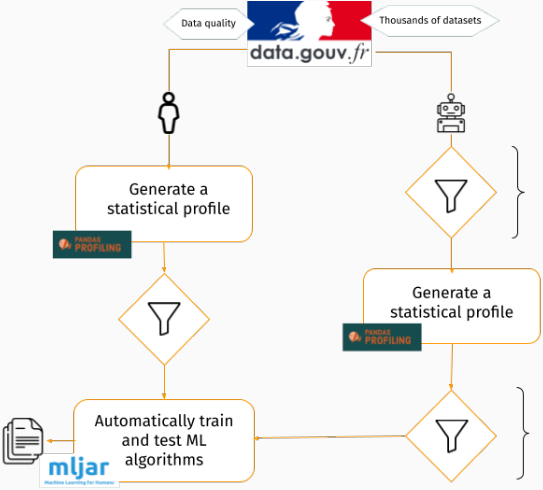
- Data format
- Data content
- Need for preprocessing

Lack of communication on Open Data platforms

Lack of catalogs specialised in Machine Learning

Our methodology for a ML data repository with Open Data

DGML: Data Gouv for Machine Learning



Data quality
data.gouv.fr
Thousands of datasets
Generate a statistical profile
PANDAS PROFILING
Generate a statistical profile
PANDAS PROFILING
Automatically train and test ML algorithms
mljar
Machine Learning for Humans

1. 200 - 2M lines
2. 3 - 500 columns
3. Both numerical and categorical variables
4. Low percentage of missing values

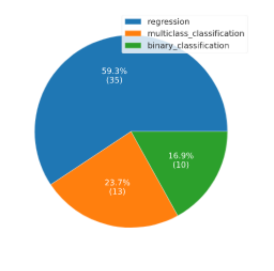
Remove:

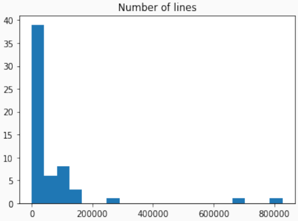
1. Categorical columns with high cardinality
2. Highly correlated columns
3. Unsupported columns

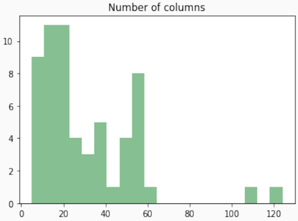
DGML: Data Gouv for Machine Learning

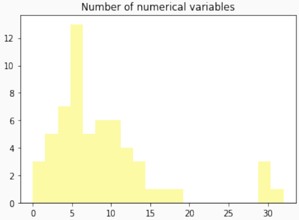
What’s in DGML?

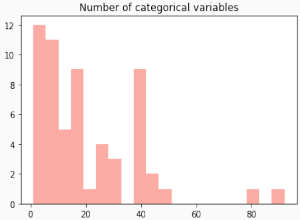
60 datasets : 10 manually selected | 50 automatically selected

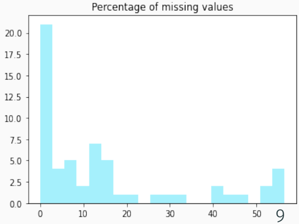



Number of lines


Number of columns

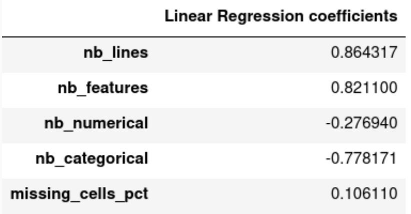

Number of numerical variables


Number of categorical variables


Percentage of missing values

The meta-features influence on the algorithms performance

- What makes a dataset a good dataset for ML ?
- Linear Regression on the meta-features of 60 datasets
- Metric value of the algorithms as target variable



	Linear Regression coefficients
nb_lines	0.864317
nb_features	0.821100
nb_numerical	-0.276940
nb_categorical	-0.778171
missing_cells_pct	0.106110

What's next?

Keep on investigating what makes a dataset a good dataset for ML

Test existing ML applications (such as scikit-learn examples) on Open Data

Increase the number of datasets on DGML

Create a stronger link with the data.gouv.fr community

Generalize our methodology to other Open Data platforms

Key takeaways

🗝 There is a **Lack of Open Data in Machine Learning** applications and research

🗝 We proposed a methodology to identify datasets that are adequate for Machine Learning

🗝 We created **DGML**, Data Gouv for Machine Learning : a centralized data repository for ML with Open Data from data.gouv.fr

---

Thank you!
giulia.santarsieri@data.gouv.fr
pavel.soriano@data.gouv.fr

🌐 https://datascience.etalab.studio/dgml/
etalab-ia/DGML