Ensembles classifiers in Class Imbalance Learning

Rebecca Leygonie, Nemanja Kostadinovic

Master 2 Big Data & Fouille de Données, Université Paris 8

14 janvier 2021

Plan

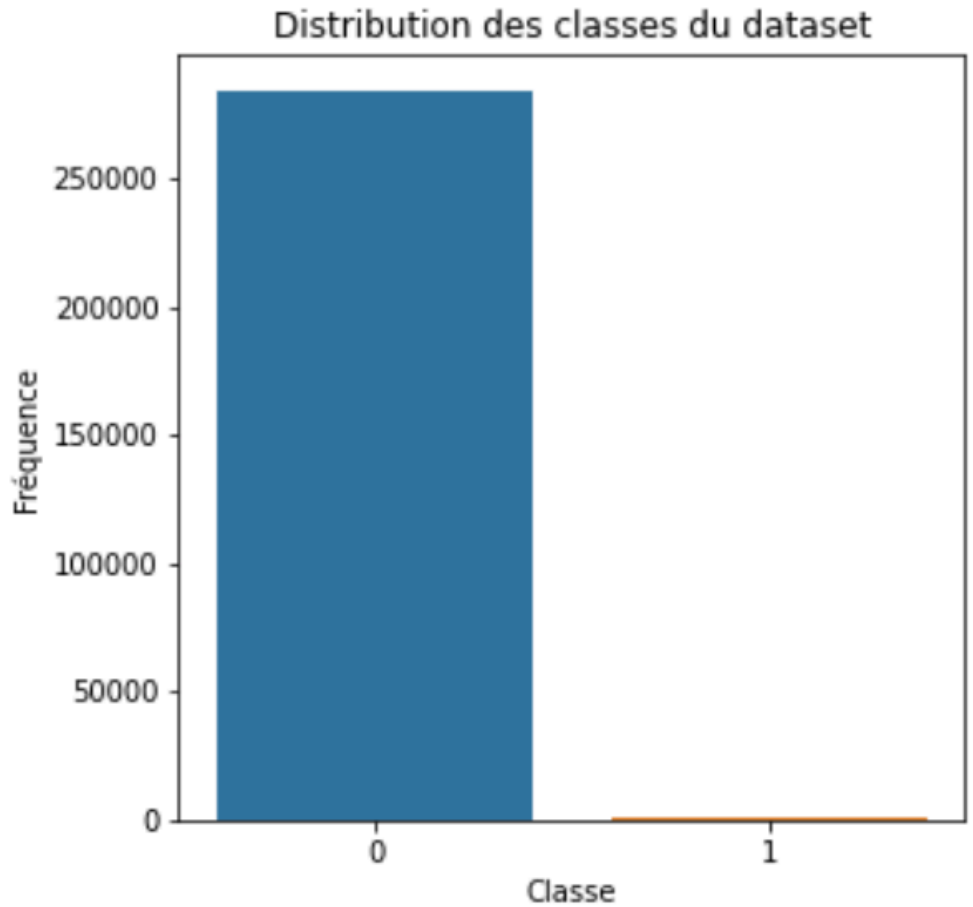
Le problème des données déséquilibrées

Données

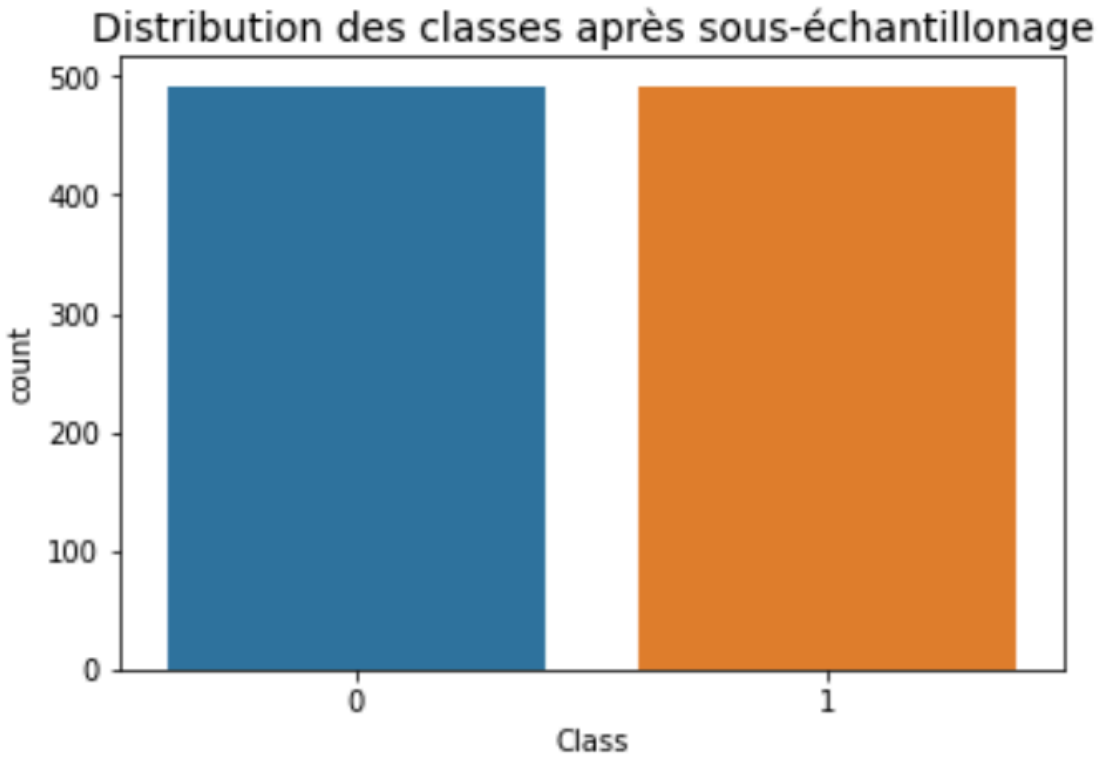
:

	Time	V2	V3	V4	V5	V6	V7	V8	Amount	Class
0	0.0	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	149.62	0
1	0.0	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	2.69	0
2	1.0	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	378.66	0
3	1.0	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	123.50	0
4	2.0	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	69.99	0
...	...	...	...	...	...	...	...	...	...	...
284802	172786.0	10.071785	-9.834783	-2.066656	-5.364473	-2.606837	-4.918215	7.305334	0.77	0
284803	172787.0	-0.055080	2.035030	-0.738589	0.868229	1.058415	0.024330	0.294869	24.79	0
284804	172788.0	-0.301254	-3.249640	-0.557828	2.630515	3.031260	-0.296827	0.708417	67.88	0
284805	172788.0	0.530483	0.702510	0.689799	-0.377961	0.623708	-0.686180	0.679145	10.00	0
284806	172792.0	-0.189733	0.703337	-0.506271	-0.012546	-0.649617	1.577006	-0.414650	217.00	0

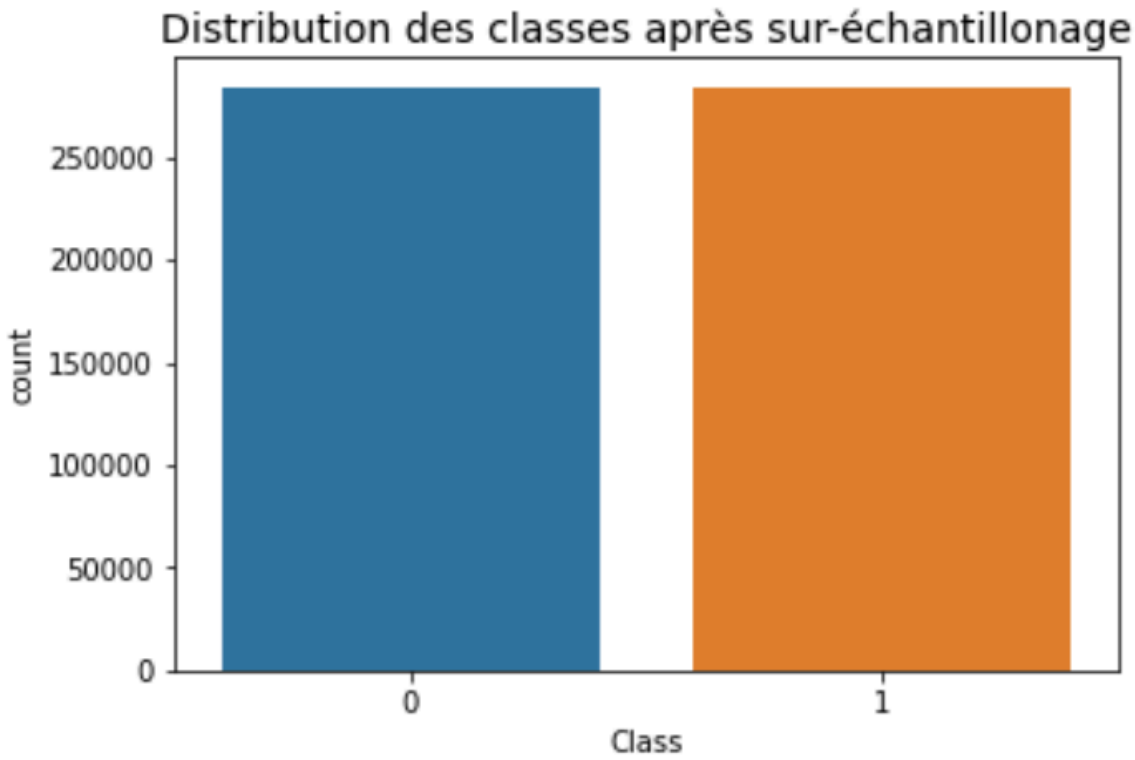
Données


Distribution des classes du dataset
250000
200000
150000
100000
50000
0
Fréquence
0
1
Classe

Sous-échantillonnage



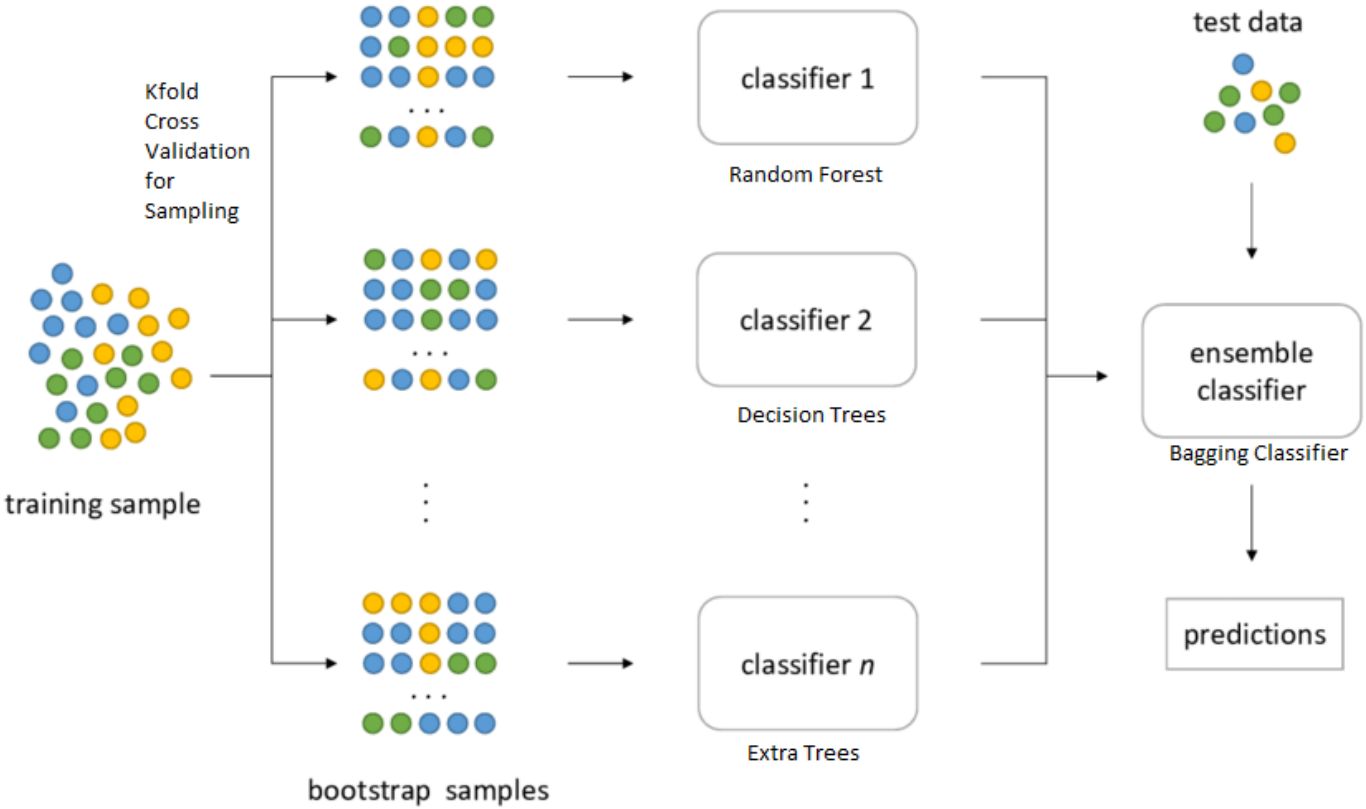
Sur-échantillonnage



Modèles

- Bagging
  - Bagging Classifier
  - Random Forest Classifier
  - Extra Trees Classifier
  - Logistic Regression
  - Gaussian Naïve Bayes
  - One Class SVM
- Boosting
  - Ada Boost Classifier
  - Gradient Boosting Classifier

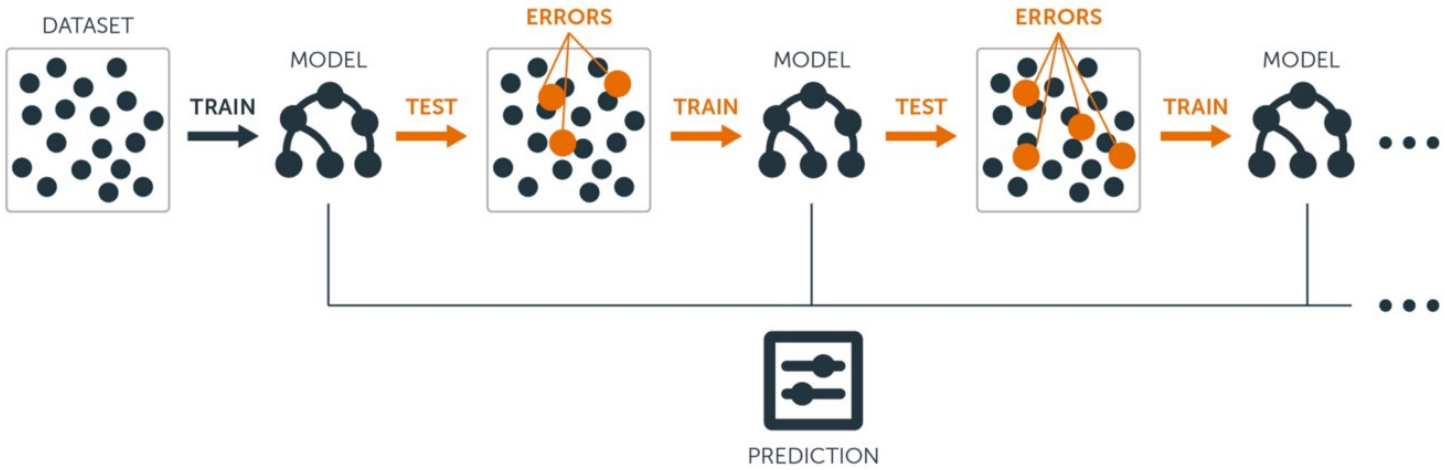
Modèles de bagging



Bagging Classifier Process Flow

Source :https ://medium.com/ml-research-lab/bagging-ensemble-meta-algorithm-for-reducing-variance-c98fffa5489f

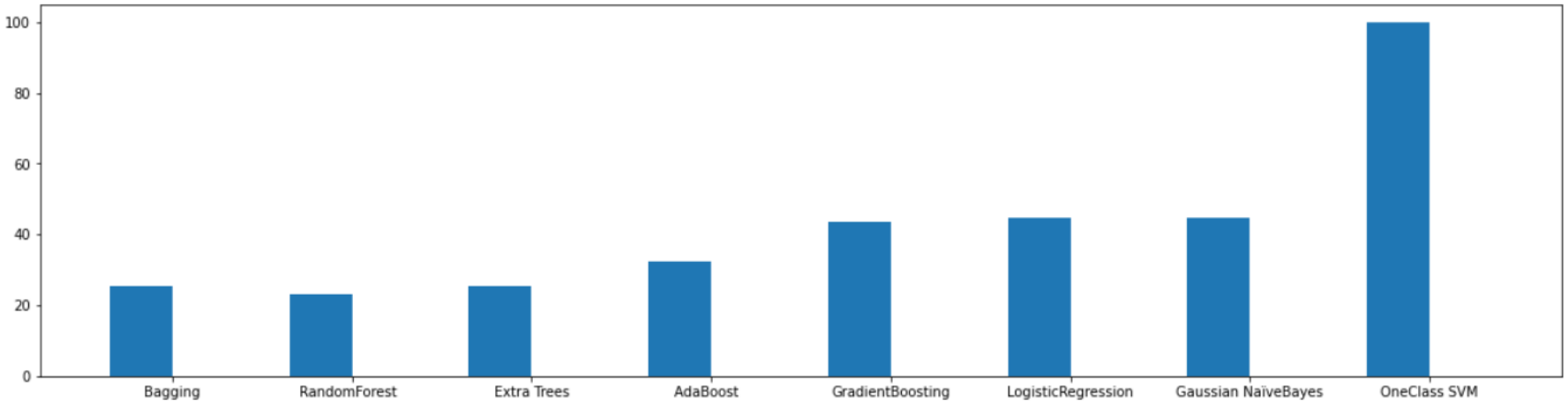
Modèles de boosting



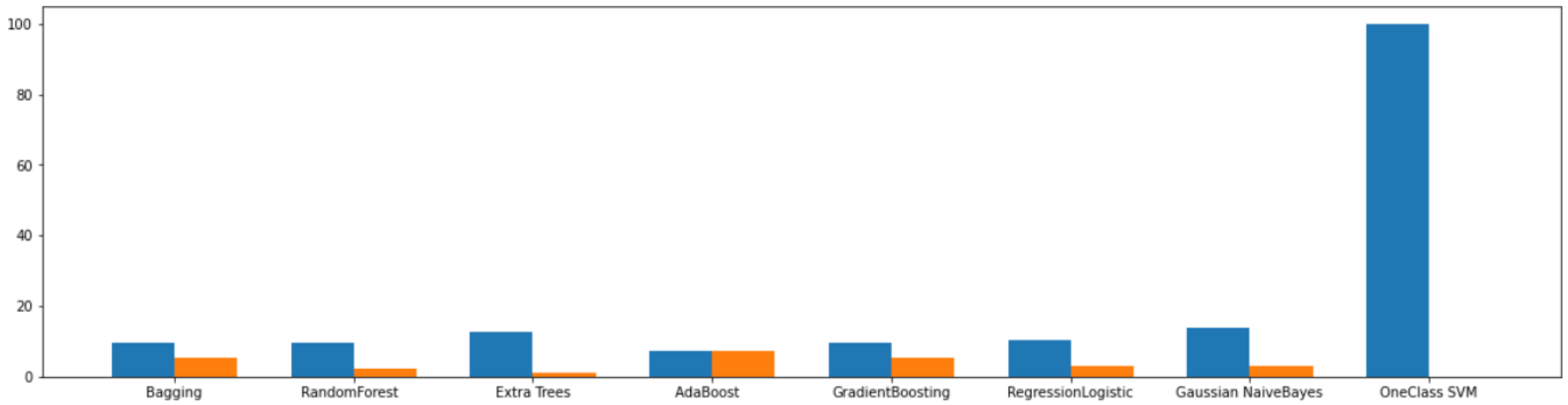
Source :https ://docs.paperspace.com/machine-learning/wiki/gradient-boosting

Application des modèles

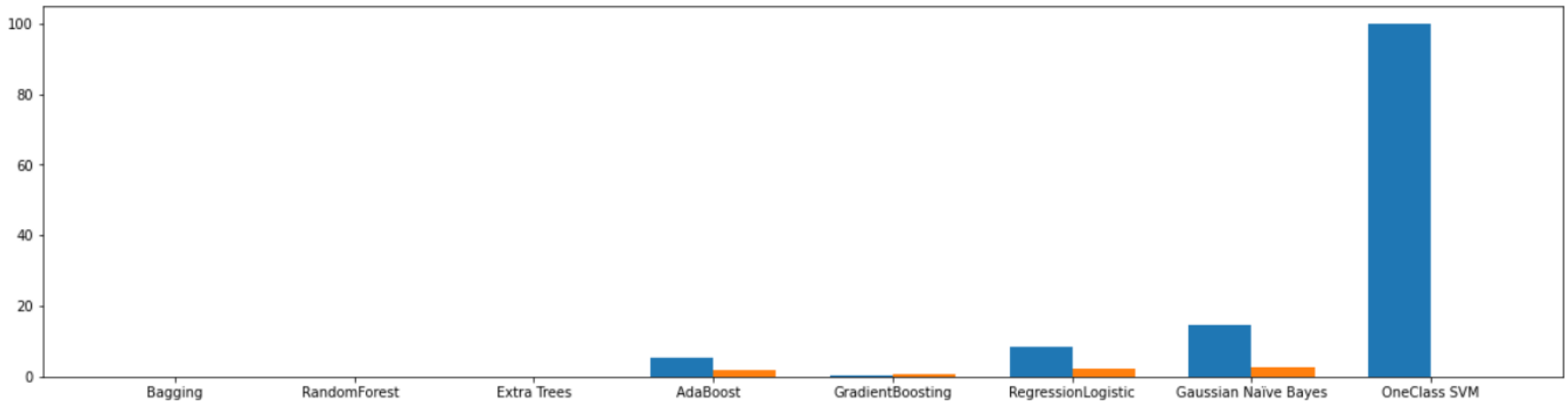
Données d'origine



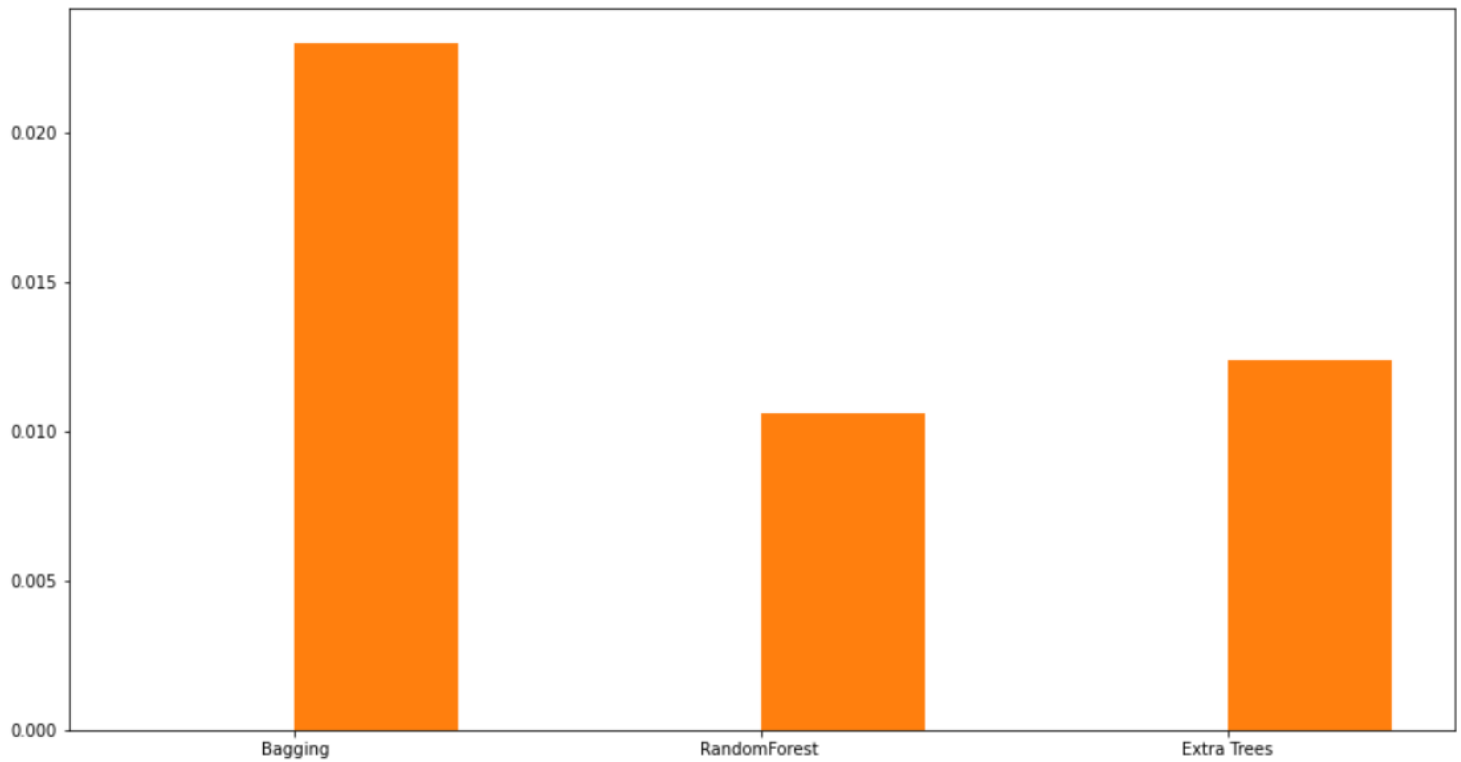
Données sous-échantillonnées



Données sur-échantillonnées



Données sur-échantillonnées



Conclusion de l'analyse

Bilan personnel

- Sukarna Barua et al. « MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning ». In : *IEEE Transactions on Knowledge and Data Engineering* 26.2 (2012), p. 405-425.
- Eric Bauer et Ron Kohavi. « An empirical comparison of voting classification algorithms : Bagging, boosting, and variants ». In : *Machine learning* 36.1-2 (1999), p. 105-139.
- Leo Breiman. « Bagging predictors ». In : *Machine learning* 24.2 (1996), p. 123-140.
- Leo Breiman. « Pasting small votes for classification in large databases and on-line ». In : *Machine learning* 36.1-2 (1999), p. 85-103.
- Yoav Freund et Robert E Schapire. « A decision-theoretic generalization of on-line learning and an application to boosting ». In : *Journal of computer and system sciences* 55.1 (1997), p. 119-139.
- Jerome H Friedman. « Greedy function approximation : a gradient boosting machine ». In : *Annals of statistics* (2001), p. 1189-1232.
- Mikel Galar et al. « A review on ensembles for the class imbalance problem : bagging-, boosting-, and hybrid-based approaches ». In : *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4 (2011), p. 463-484.
- Pierre Geurts, Damien Ernst et Louis Wehenkel. « Extremely randomized trees ». In : *Machine learning* 63.1 (2006), p. 3-42.
- Tin Kam Ho. « Random decision forests ». In : *Proceedings of 3rd international conference on document analysis and recognition*. T. 1. IEEE. 1995, p. 278-282.
- Tin Kam Ho. « The random subspace method for constructing decision forests ». In : *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998), p. 832-844.

- Xu-Ying Liu, Jianxin Wu et Zhi-Hua Zhou. « Exploratory undersampling for class-imbalance learning ». In : *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2008), p. 539-550.
- D. Opitz et R. Maclin. « Popular Ensemble Methods : An Empirical Study ». In : *Journal of Artificial Intelligence Research* 11 (août 1999), p. 169-198. ISSN : 1076-9757. DOI : 10.1613/jair.614. URL : http://dx.doi.org/10.1613/jair.614.
- Bernhard Schölkopf et al. « Support Vector Method for Novelty Detection ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de S. Solla, T. Leen et K. Müller. T. 12. MIT Press, 2000, p. 582-588. URL : https://proceedings.neurips.cc/paper/1999/file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf.
- Harry Zhang. « The Optimality of Naive Bayes ». In : *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. May 17-19, 2004 (Miami Beach, Florida, USA). Sous la dir. de Valerie Barr et Zdravko Markov. AAAI Press, 2004.