Rapport Machine Learning :

Auteurs : Cornen Paul, Fouquet Ulysse, Collet William, Guermeur Nicolas & Léon Pierre

Modèle	Tâche	RMSE / Accuracy	R^2 / F1	Temps d'entraînement (fit)	Temps de prédiction (pred)
DecisionTree	Régression	**14.1654**	**0.5719**	37.2 ms	0.3 ms
DecisionTree (scikit)	Régression	**14.1654**	**0.5719**	**5.6 ms**	**0.2 ms**
DecisionTree	Classification	**0.775**	**0.7**	45.7 ms	**0.1 ms**
DecisionTree (scikit)	Classification	0.725	0.633	**2.2 ms**	0.2 ms
RandomForest	Régression	12.8886	0.6456	5946.9 ms	37.5ms
RandomForest (scikit)	Régression	**12.6482**	**0.6587**	**337.9 ms**	**36.2 ms**
RandomForest	Classification	**0.812**	**0.714**	4426.6 ms	38.0 ms
RandomForest (scikit)	Classification	0.787	0.702	**168.9 ms**	**29.4 ms**
Ridge	Régression	**13.5374**	**0.6090**	**0.3 ms**	**0.0 ms**
Ridge (scikit)	Régression	**13.5374**	**0.6090**	1.2 ms	0.1 ms
Lasso	Régression	**13.5116**	**0.6105**	19.7 ms	**0.1 ms**
Lasso (scikit)	Régression	13.5219	0.6087	**2.6 ms**	**0.1 ms**

SVM	Régression	13.6186	0.6043	**63.2 ms**	**0.0 ms**
SVM (scikit)	Régression	**12.7355**	**0.6540**	67.7 ms	29.6 ms
SVM	Classification	**0.875**	**0.852**	5.2 ms	**0.0 ms**
SVM (scikit)	Classification	0.850	0.821	**5.1 ms**	0.1 ms

Métriques : R^2, RMSE pour la régression et F1, Accuracy pour la classification

Dans la partie traitement des données, on a décidé de garder “maxO3v” (si on veut l'enlever il n'y a qu'à décommenter la ligne 62 de “utils.py”). Avant de voir pourquoi on a pris cette décision, il faut d'abord déterminer ce qu'est “maxO3v”. En regardant les données, il faut noter 3 points importants: “id”, “maxO3”, et “maxO3v”. “id” est la date de l'enregistrement au format YYYYMMDD. ”maxO3”, et “maxO3v” sont des valeurs en point flottant. En regardant chaque ligne on peut voir que “maxO3” et “maxO3v” se ressemblent beaucoup, à un détail près: “maxO3v” est décalé. Chaque valeur de “maxO3” se répète dans “maxO3v” le lendemain. En d'autres mots, les valeurs de “maxO3v” sont celles de “maxO3” de la **v**eille. Ça permet au modèle de savoir que les valeurs ne seront pas trop éloignées de celles de la veille. Mais cette information supplémentaire ne permet pas au modèle de "tricher" en utilisant le “maxO3v” du lendemain comme un humain pourrait le faire.

Conclusion

Dans l’ensemble, les modèles développés manuellement atteignent des scores de performance similaires à ceux des implémentations scikit-learn, confirmant la validité des approches. Cependant, les temps d’entraînement et de prédiction sont systématiquement plus longs, en raison de l’absence d’optimisations internes (vectorisation, parallélisation). Les différences sont particulièrement marquées pour les algorithmes complexes comme les forêts aléatoires, tandis que les modèles linéaires restent proches en vitesse et en précision.