

PRÉVISIONS MÉTÉO EN AUSTRALIE

PROJET DATA SCIENCE
BOOTCAMP NOV 2024

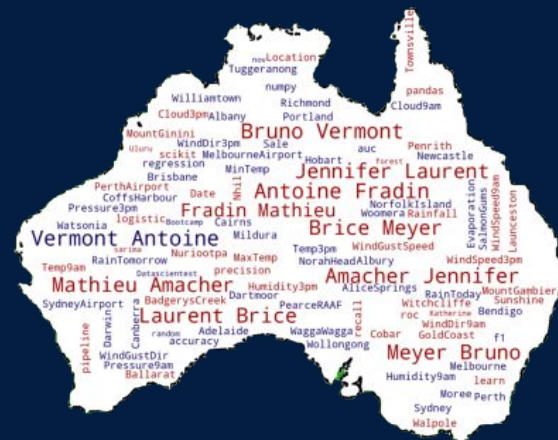
Présenté par :

Mathieu Amacher

Jennifer Laurent

Brice Meyer

Bruno Vermont



Mentor : Antoine Fradin





INTRODUCTION

CONTEXTE

- Stations météorologiques à travers l'Australie
- Informations météorologiques quotidiennes sur 10 ans

OBJECTIFS

- Entraîner des modèles de machine learning pour les prévisions météorologiques
- Prédire la probabilité de pluie
- Prédire les autres variables météorologiques types (températures, l'humidité ...)





SOMMAIRE

1 Présentation du jeu de données

2 Pré-processing

3 Modélisations

3.1 Probabilité de pluie

3.2 Probabilité de pluie par station

3.3 Prédictions des autres variables

4 Conclusion

1

PRÉSENTATION DU JEU DE DONNÉES

kaggle

Présentation Globale

Novembre 2007 - Juin 2017 (sauf avril-mai 2011)
145 460 données, 14 quantitatives + 9 qualitatives
49 Stations : 3 000 données par station

Variables Ponctuelles (9am et 3pm)

Quantitatives : WindSpeed, Humidity, Pressure, Temp
Qualitatives : WindDir, Cloud

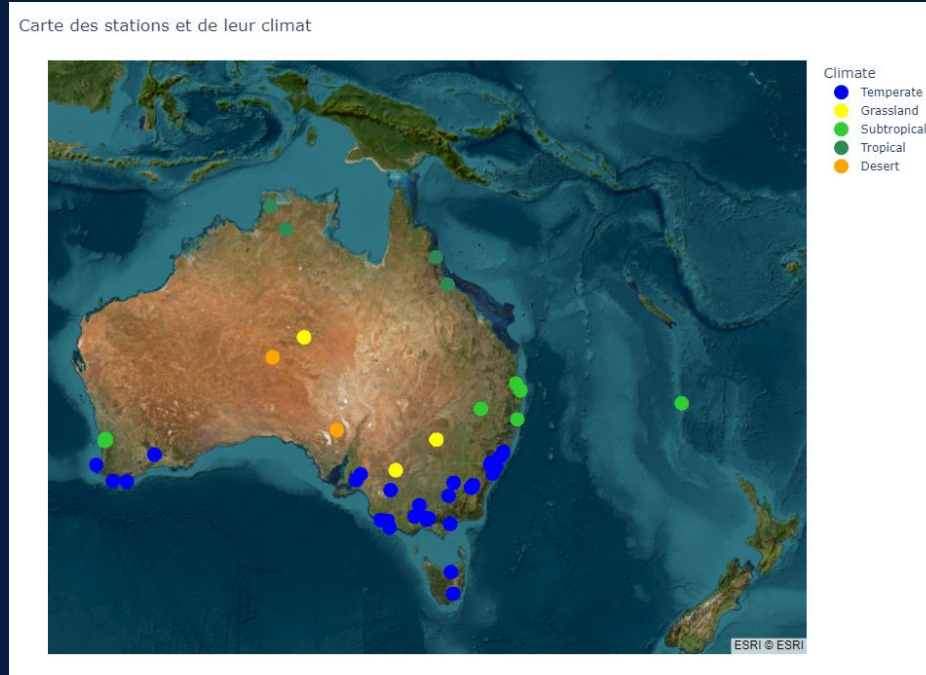
Variables Globales

Quantitatives : MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustSpeed
Qualitatives : Date, Location, WindGustDir, RainToday, **RainTomorrow**

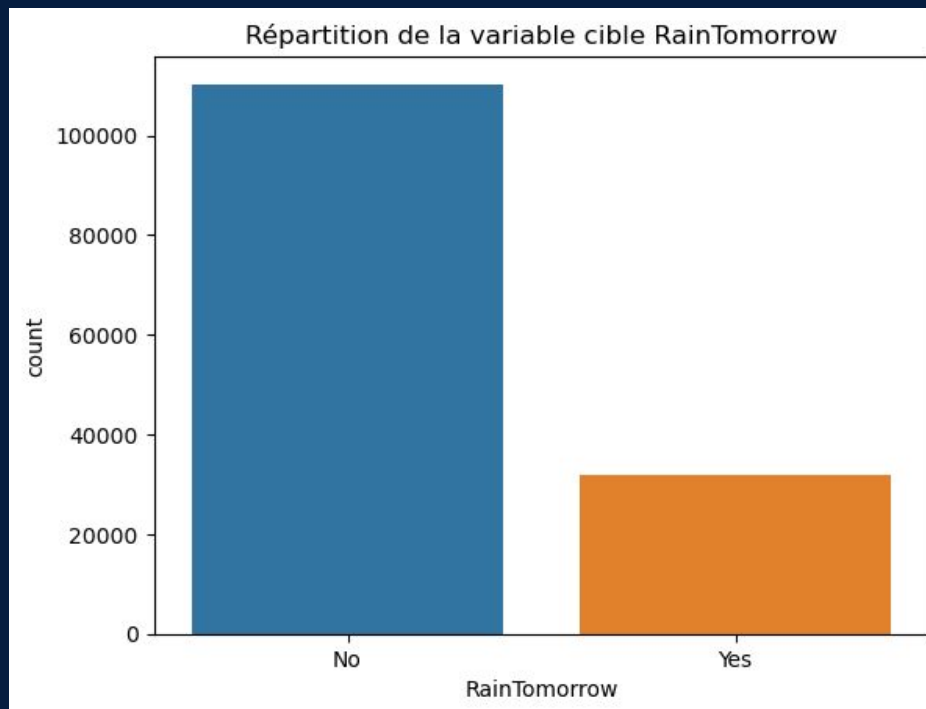


1

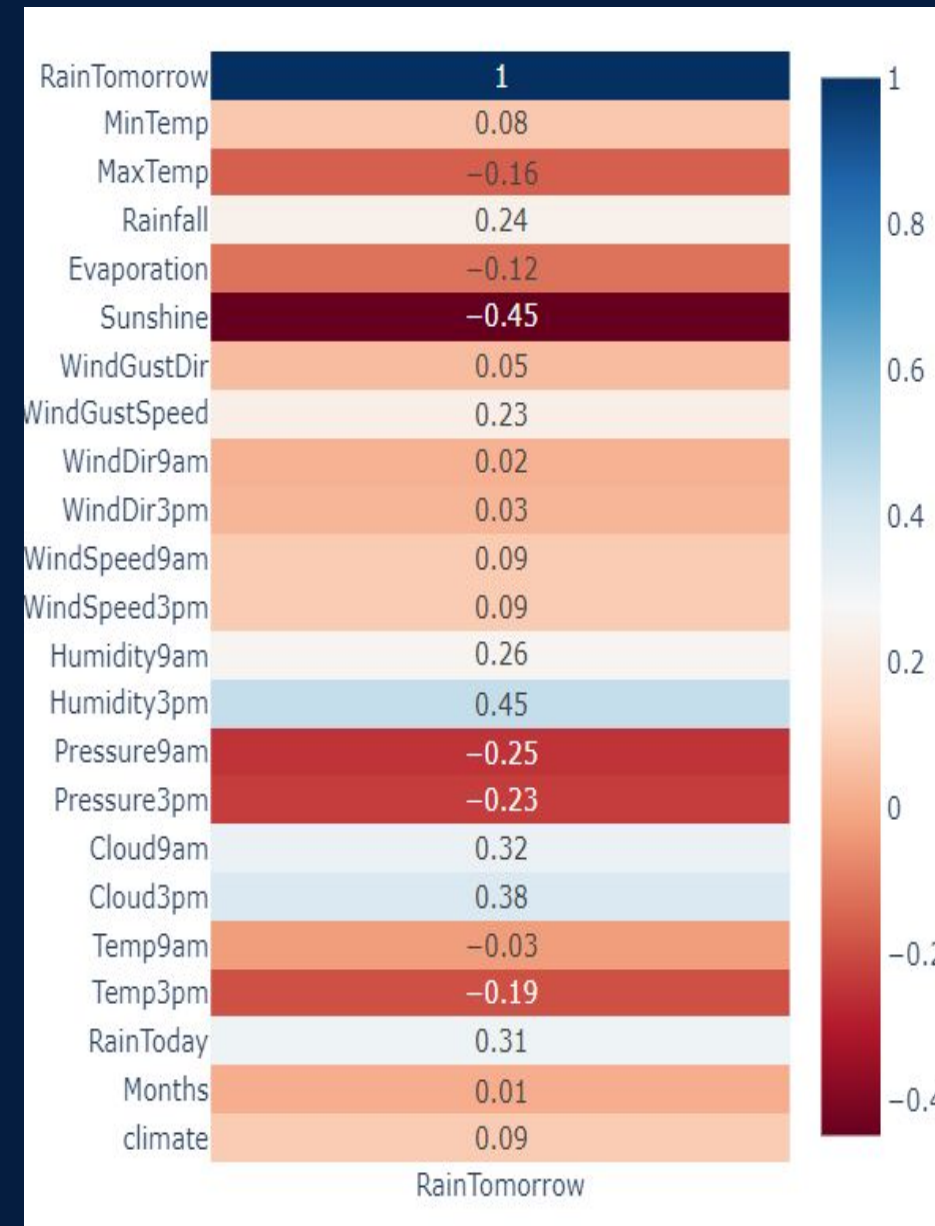
PRÉSENTATION DU JEU DE DONNÉES



Stations majoritairement au sud
32/49 dans climat tempéré



Fort déséquilibre de la variable cible



Sunshine, Humidity3pm, Cloud3pm
les plus corrélées à RainTomorrow

Présence de NAs en %

Date	0.00
Location	0.00
MinTemp	1.02
MaxTemp	0.87
Rainfall	2.24
Evaporation	43.17
Sunshine	48.01
WindGustDir	7.10
WindGustSpeed	7.06
WindDir9am	7.26
WindDir3pm	2.91
WindSpeed9am	1.21
WindSpeed3pm	2.11
Humidity9am	1.82
Humidity3pm	3.10
Pressure9am	10.36
Pressure3pm	10.33
Cloud9am	38.42
Cloud3pm	40.81
Temp9am	1.21
Temp3pm	2.48
RainToday	2.24
RainTomorrow	2.25

Peu de NAs pour l'ensemble des variables
sauf Sunshine, Evaporation, Cloud :
Fort taux de NA

2 PRÉ-PROCESSING

Gestion des NA

Complétion NA variable cible :

- RainToday jour suivant
- Quantité de Rainfall jour suivant

Preprocessing itératif : ajustements après modélisations

Etape 1 :

Test de différentes méthodes

→ Médiane/mode selon location et mois

Etape 2 :

- Station la plus proche (50 kms)
- Médiane/mode selon station et mois
- Médiane/Mode selon station et valeur variable cible



2 PRÉ-PROCESSING

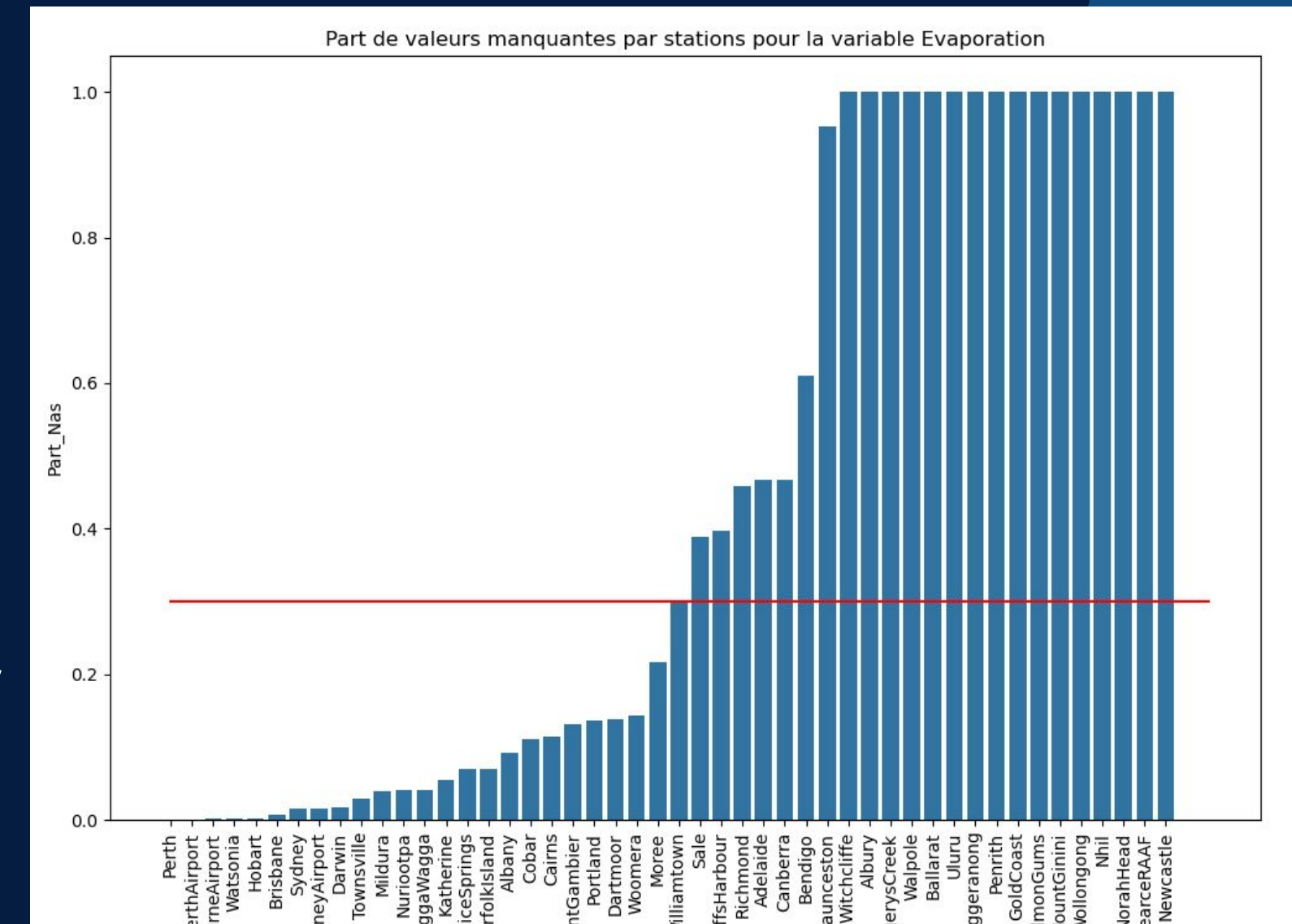
Gestion des NA

Etape 3 : Enrichissement du dataset

- Ajout de variables lagguées
- Ajout de moyennes glissantes
- Ajout du taux de variation entre 9am et 3pm
- Gestion de la multicollinéarité

Etape 4 :

- Preprocessing par station et pas sur le dataset entier
- Réintégration de certaines variables supprimées
→ si -30% valeurs manquantes.



3

STRATEGIE DE MODELISATION

► **Modèle “global”**



► **Modèles par
station**



► **Modélisation
temporelle**

Variable cible : RainTomorrow

Classification binaire

Preprocessing global

Modèle entraîné sur
toutes les données

Variable cible : RainTomorrow

Classification binaire

Preprocessing et entraînement
par station

Variables cibles : autres
variables météo

Série temporelle

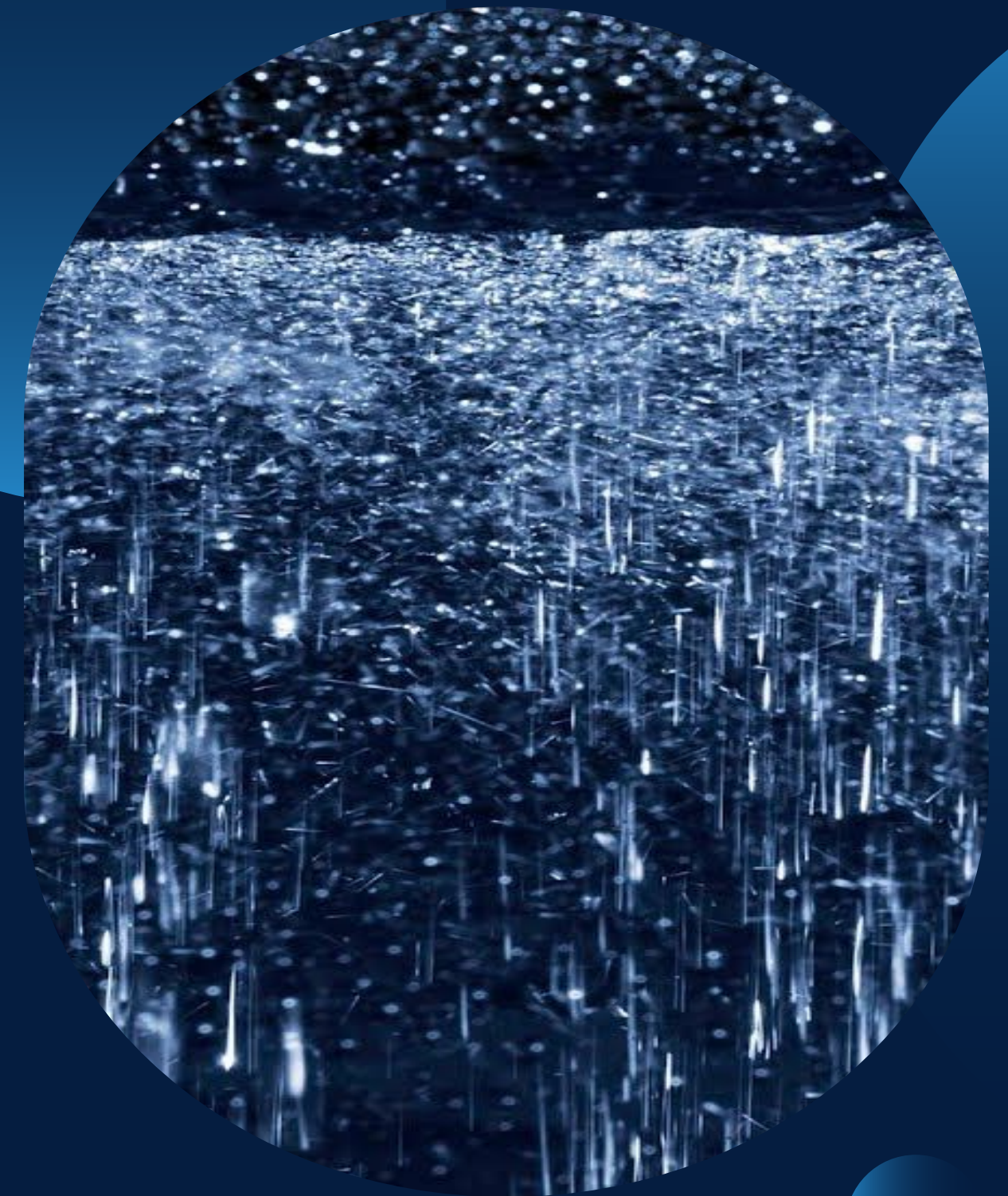
Estimation par station

3.1 MODÉLISATION DE LA PROBABILITÉ DE PLUIE

Variable cible : RainTomorrow

Ensemble du jeu de données

Démo  Streamlit [MeteoStralia](#)



3.1 MODÈLE GLOBAL : Résultats généraux

Plafond accuracy (0.86) et f1-score (0.64)

“Meilleurs modèles équilibrés” pour prédire la classe positive

Précision 0.77 : RandomForest, gridsearch basée sur le f0.5_score

Rappel 0.76 : BalancedRandomForest, gridsearch basée sur le f1_score

Gestion de la multicolinéarité

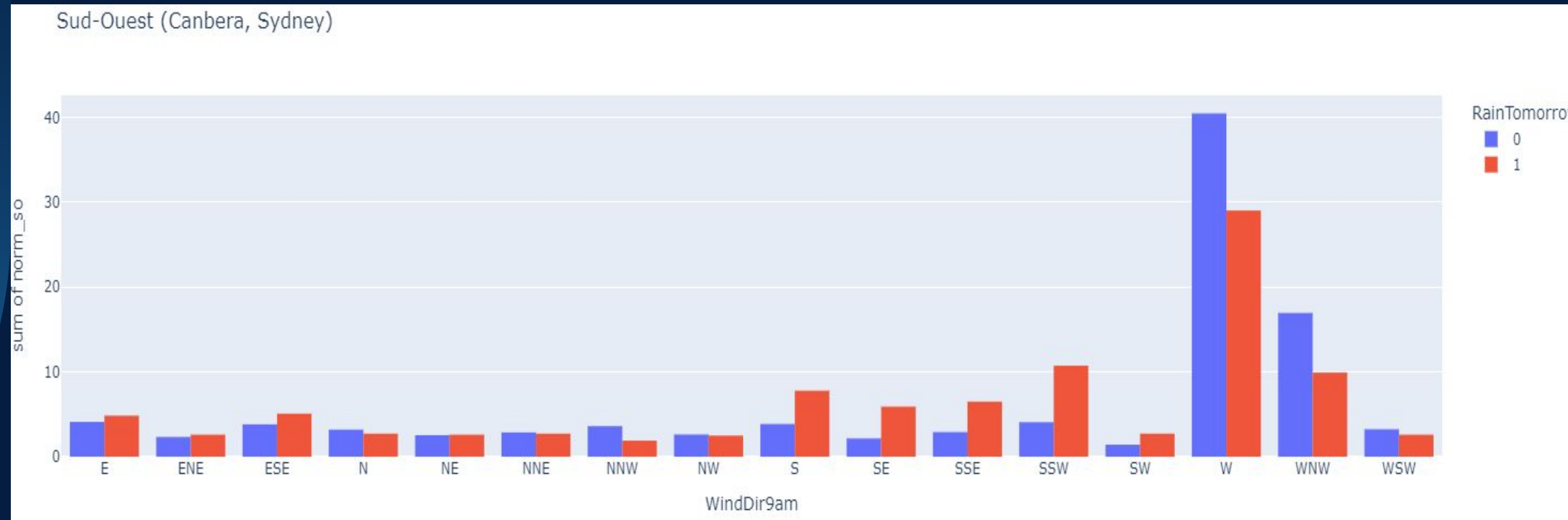
→ Meilleure interprétabilité mais moins bonnes performances

Performances variées selon la station

→ ex : RandomForest → Entre 1% et 10% de faux positifs par station

3.2 MODÉLISATION PAR STATION

Le Biais Spatial



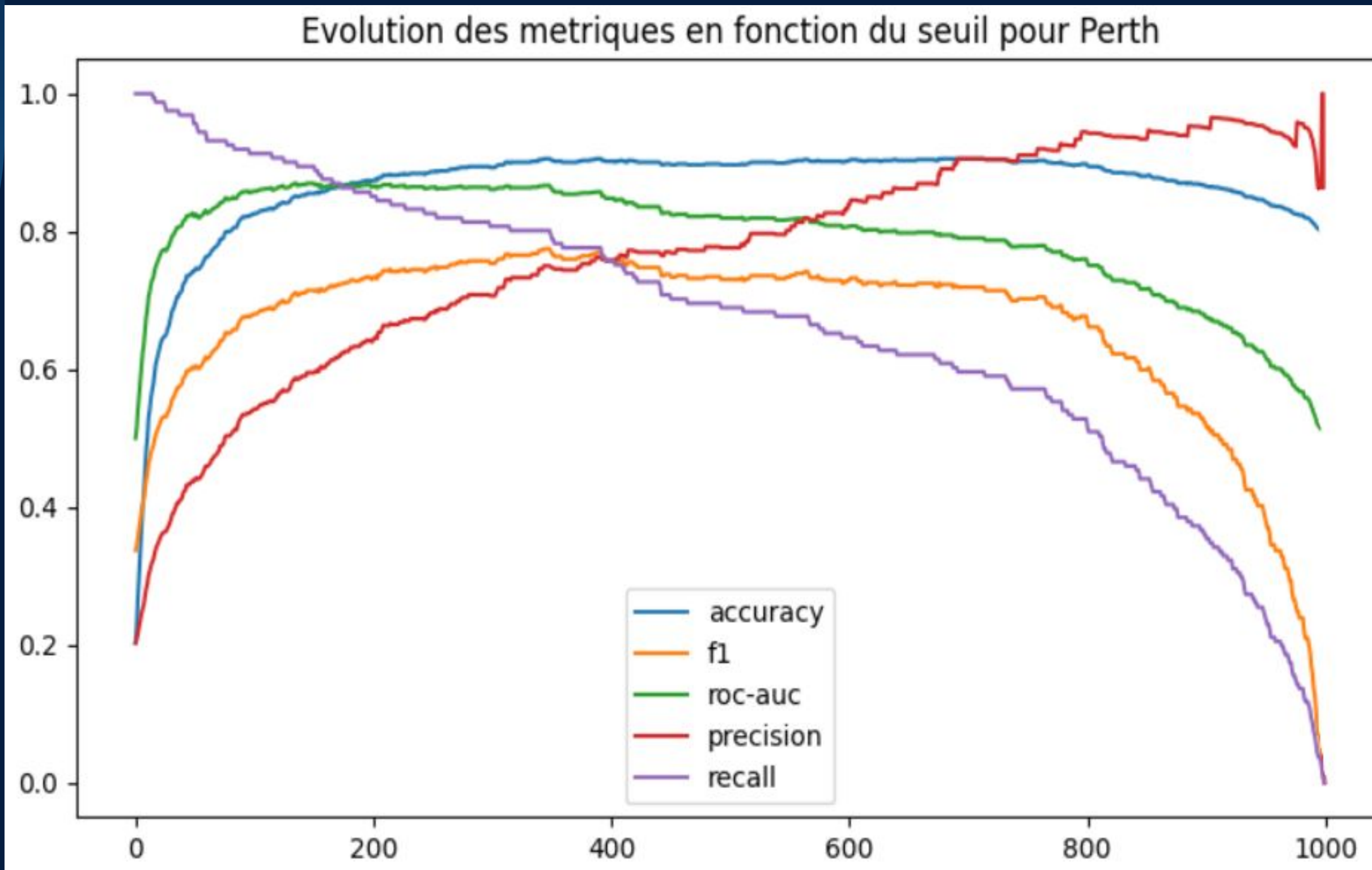
Pour les territoires du Sud-ouest, la majorité des vents vient d'Ouest



Pour les territoires du Centre, la majorité des vents vient d'Est

3.2 MODÉLISATION PAR STATION

Seuil de Décision



Quelle métrique privilégier ?

Quelle problématique pour le consommateur final ?

cf. annexe A.5

3.2 MODÉLISATION PAR STATION

Ajout de données

	precision	recall	f1-score	support
0.0	0.71	0.98	0.82	340
1.0	0.81	0.20	0.32	169
accuracy			0.72	509
macro avg	0.76	0.59	0.57	509
weighted avg	0.74	0.72	0.66	509

Coffs Harbour

	precision	recall	f1-score	support
0.0	0.71	0.99	0.83	357
1.0	0.94	0.24	0.39	192
accuracy			0.73	549
macro avg	0.82	0.62	0.61	549
weighted avg	0.79	0.73	0.67	549

classe 0 : 65 % / classe 1 : 35 %

+8 % d'occurrence

Accuracy : + 1.4 %
Précision : + 16 %
Rappel : + 20 %
F1-score : + 21 %

MeteoStralia

3.3

MODÉLISATION DES AUTRES VARIABLES MÉTÉO

MODÉLISATIONS ET PRÉDICTIONS

Variables utilisées pour prédire RainTomorrow

MODÈLE SARIMA

Séries temporelles

Long terme

Variations saisonnières

CAS D'ÉTUDE

Sydney

Variables Wind

Démo  Streamlit

[MeteoStralia](#)

3.3 MODÉLISATION DES VARIABLES MÉTÉO

Évaluation du modèle pour Sydney

R2 élevé

MaxTemp, MinTemp, Humidity9am,
Humidity3pm, Pressure9am, Pressure3pm

R2 moyen

Rainfall, Evaporation, Sunshine, Cloud9am,
Cloud3pm

R2 faible

Wind...

Location	Variable	R2	MSE	RMSE	MAE
Sydney	MaxTemp	0,94	1,29	1,14	0,81
Sydney	MinTemp	0,87	2,65	1,63	1,36
Sydney	Rainfall	0,36	80,71	8,98	4,24
Sydney	Evaporation	0,54	4,06	2,02	1,54
Sydney	Sunshine	0,45	7,84	2,80	2,29
Sydney	WindGustSpeed	-0,30	203,77	14,28	12,17
Sydney	Humidity9am	0,75	53,69	7,33	5,73
Sydney	Humidity3pm	0,82	43,80	6,62	5,21
Sydney	Pressure9am	0,96	1,95	1,40	1,07
Sydney	Pressure3pm	0,96	2,12	1,46	1,08
Sydney	Cloud9am	0,46	3,98	2,00	1,65
Sydney	Cloud3pm	0,35	4,49	2,12	1,76
Sydney	WindGustDir_cos	-0,03	0,60	0,78	0,66
Sydney	WindGustDir_sin	-0,03	0,42	0,65	0,52
Sydney	WindDir9am_cos	0,20	0,43	0,66	0,55
Sydney	WindDir9am_sin	0,10	0,33	0,58	0,49
Sydney	WindDir3pm_cos	0,10	0,43	0,65	0,57
Sydney	WindDir3pm_sin	0,10	0,44	0,67	0,58

3.3 MODÉLISATION DES VARIABLES MÉTÉO

Variables Wind

- Vitesse et direction du vent
- Performances du modèle varient considérablement
- Nature complexe et chaotique du vent
- Dépend de facteurs qui peuvent varier rapidement et de manière non linéaire
- Encodées en utilisant des fonctions trigonométriques (cosinus et sinus)

Location	Variable	R2	MSE	RMSE	MAE
Sydney	WindGustSpeed	-0,30	203,77	14,28	12,17
Adelaide	WindGustSpeed	0,05	112,56	10,61	8,74
AliceSprings	WindGustSpeed	0,48	52,42	7,24	5,34
Brisbane	WindGustSpeed	0,45	33,74	5,81	4,22
Cairns	WindGustSpeed	0,67	23,65	4,86	3,73
Canberra	WindGustSpeed	0,58	88,12	9,39	7,26
Darwin	WindGustSpeed	0,03	101,37	10,07	7,99
Hobart	WindGustSpeed	0,67	106,96	10,34	7,94
Melbourne	WindGustSpeed	0,38	113,32	10,65	8,76
Perth	WindGustSpeed	0,40	47,51	6,89	5,43
Uluru	WindGustSpeed	0,24	78,99	8,89	7,22
Sydney	WindGustDir_cos	-0,03	0,60	0,78	0,66
Adelaide	WindGustDir_cos	-3,51	2,23	1,49	1,26
AliceSprings	WindGustDir_cos	-0,08	0,33	0,57	0,46
Brisbane	WindGustDir_cos	0,12	0,42	0,65	0,57
Cairns	WindGustDir_cos	-0,85	0,48	0,69	0,59
Canberra	WindGustDir_cos	-0,37	0,55	0,74	0,65
Darwin	WindGustDir_cos	-0,28	0,52	0,72	0,63
Hobart	WindGustDir_cos	0,10	0,31	0,56	0,45
Melbourne	WindGustDir_cos	-0,38	0,54	0,74	0,57
Perth	WindGustDir_cos	-0,01	0,55	0,74	0,65
Uluru	WindGustDir_cos	0,07	0,43	0,65	0,57
Sydney	WindGustDir_sin	-0,03	0,42	0,65	0,52
Adelaide	WindGustDir_sin	0,08	0,40	0,63	0,54
AliceSprings	WindGustDir_sin	-0,35	0,73	0,85	0,66
Brisbane	WindGustDir_sin	0,23	0,40	0,63	0,52
Cairns	WindGustDir_sin	0,45	0,29	0,54	0,46
Canberra	WindGustDir_sin	0,05	0,56	0,75	0,66
Darwin	WindGustDir_sin	-0,24	0,72	0,85	0,70
Hobart	WindGustDir_sin	0,14	0,55	0,74	0,67
Melbourne	WindGustDir_sin	0,13	0,49	0,70	0,60
Perth	WindGustDir_sin	0,07	0,41	0,64	0,56
Uluru	WindGustDir_sin	0,05	0,45	0,67	0,54

4 CONCLUSION ET PERSPECTIVES

Probabilité de pluie pour le jeu de données complet

- Palier maximal de performance globale -> ajout de données ?
- Classifier et paramètres -> choix entre précision et recall mais difficile à affiner par station
- Ne permet pas de prendre en compte les variables disponibles localement uniquement
- Deep learning pour des données de panel?

Probabilité de pluie par station

- Amélioration possible en récupérant des données par scraping ou API payante
- Tester le modèle en supprimant des variables explicatives (ex : *direction de vent*)

Prévisions des autres variables météorologiques

- SARIMA → Performant sauf pour “Wind” ⇒ Modifier les Paramètres (ACF & PACF)
- Deep learning → SARIMA + LSTM ⇒ Capter la saisonnalité et les tendances complexes

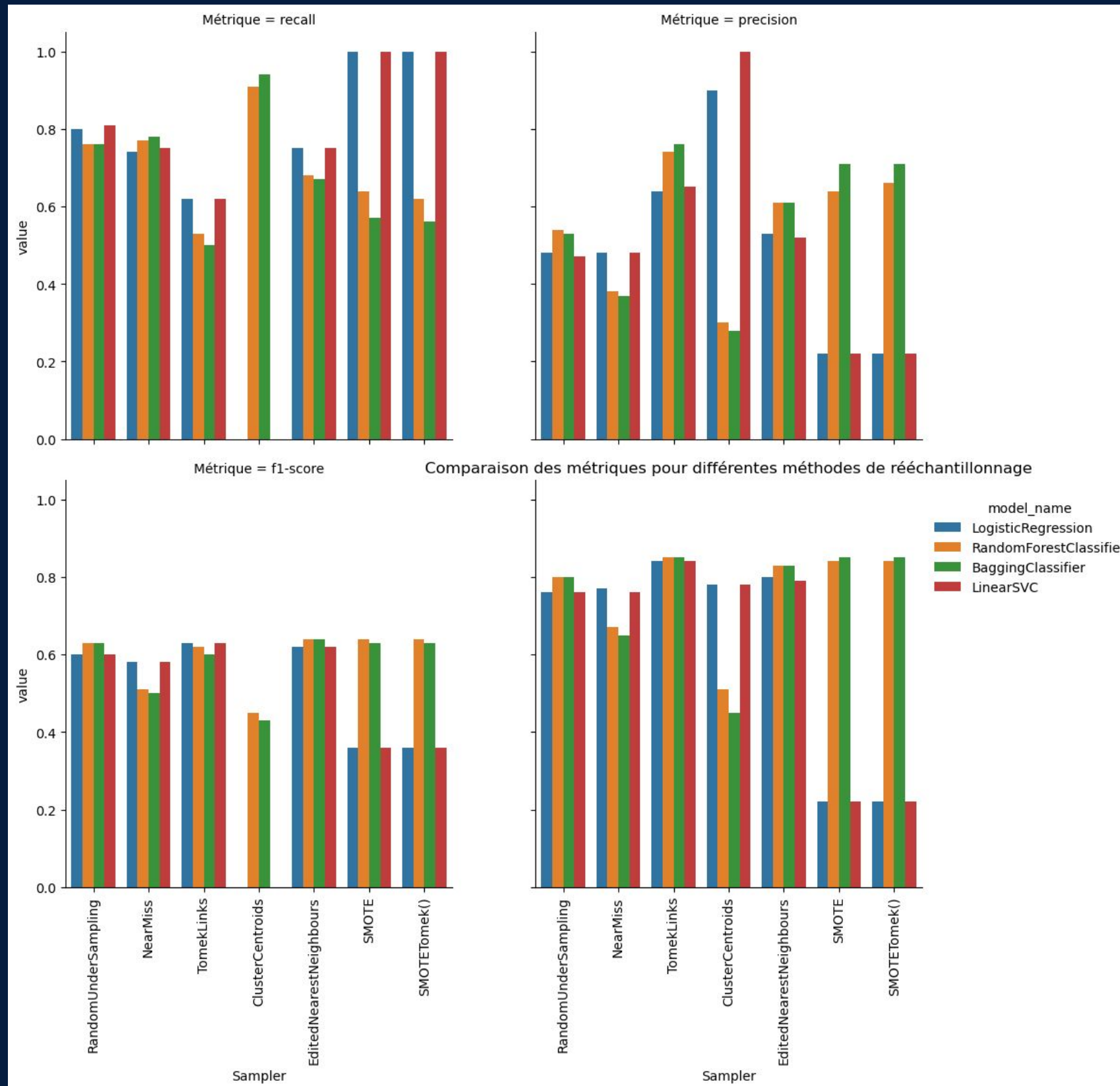
**MERCI
POUR VOTRE
ATTENTION**

Des questions ?

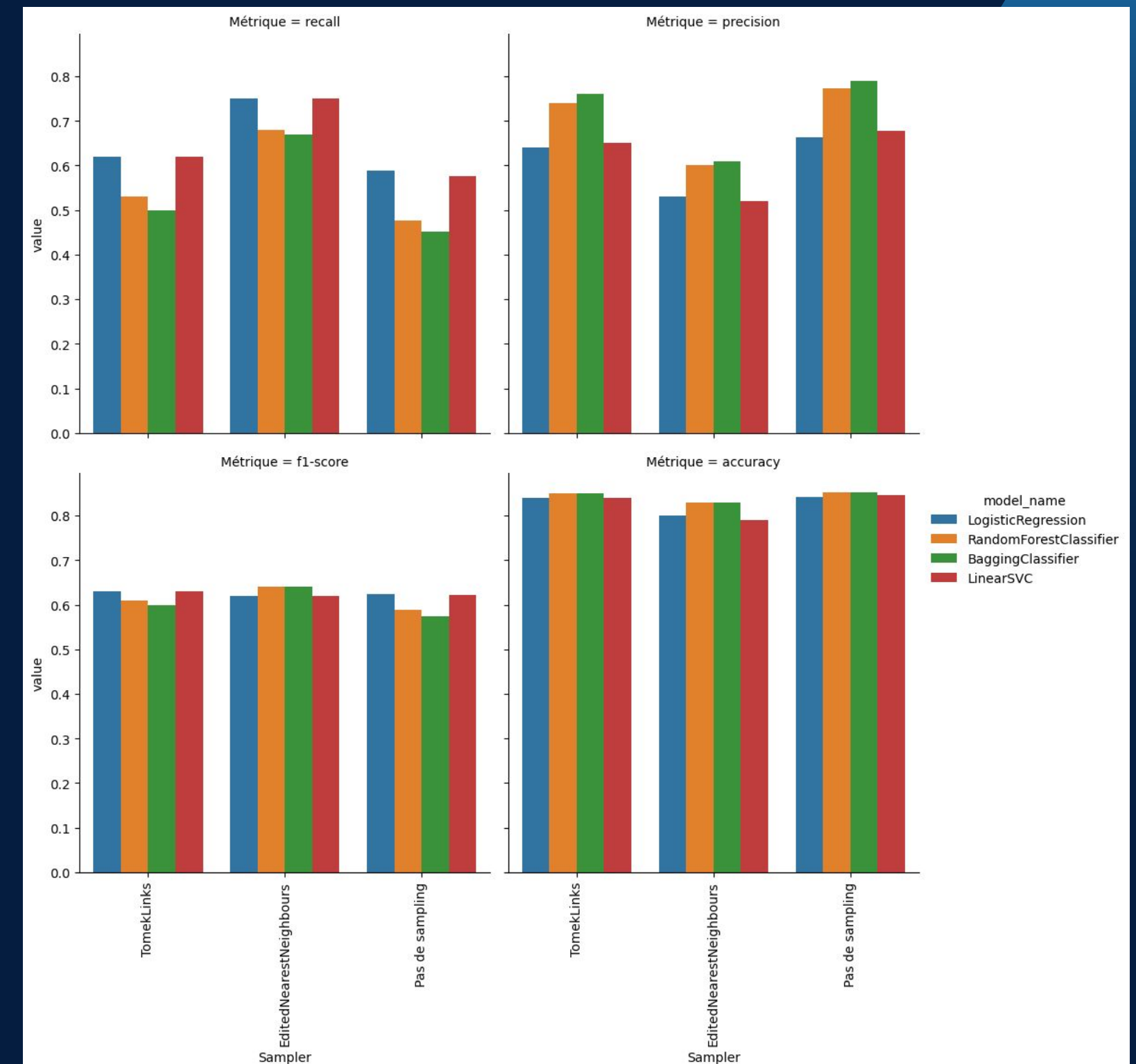


A.1 Annexes : Méthodes de gestion du déséquilibre

Comparaison des méthodes

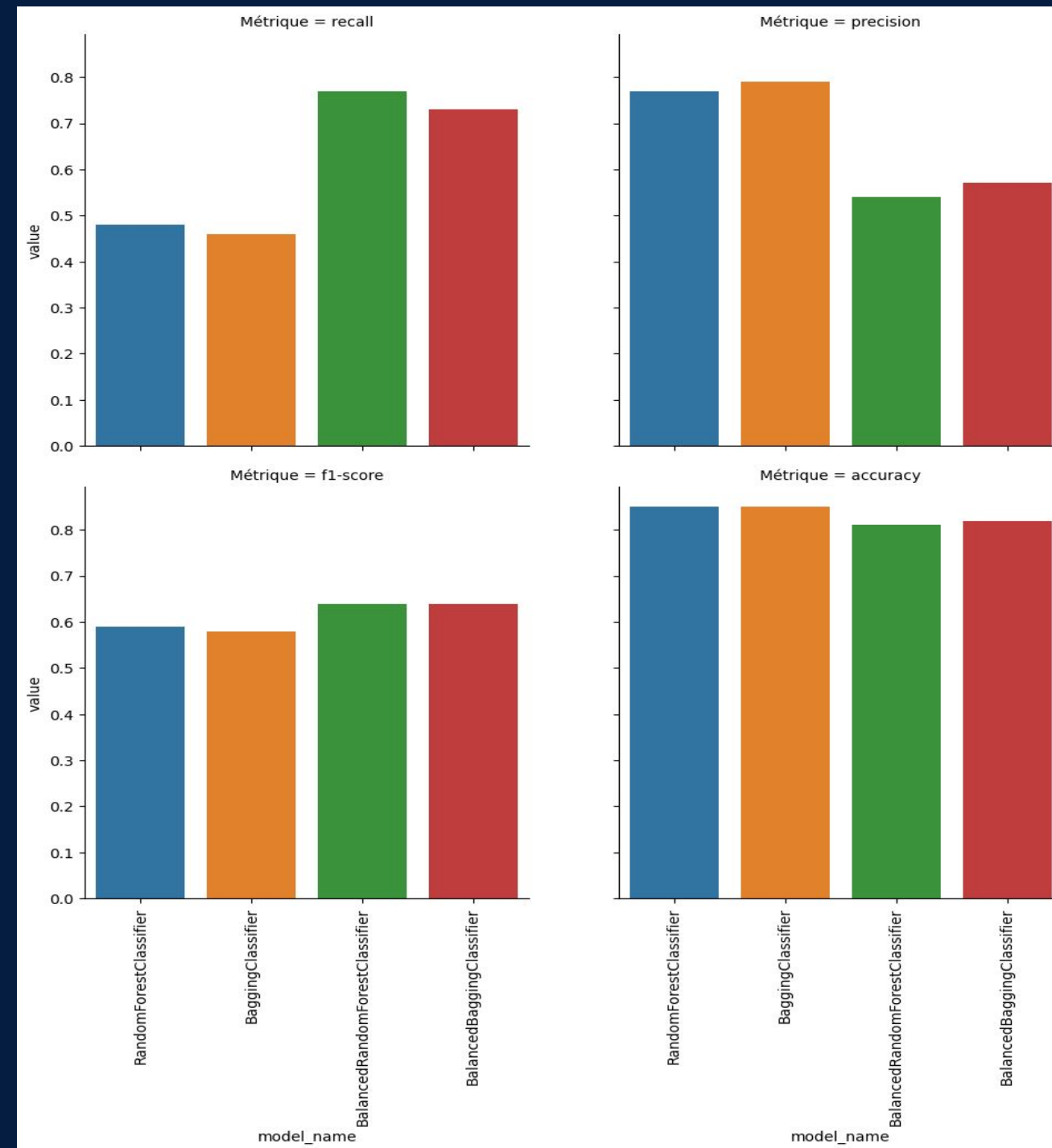


Sans rééchantillonnage VS avec



A.2 Annexes : Méthodes de gestion du déséquilibre

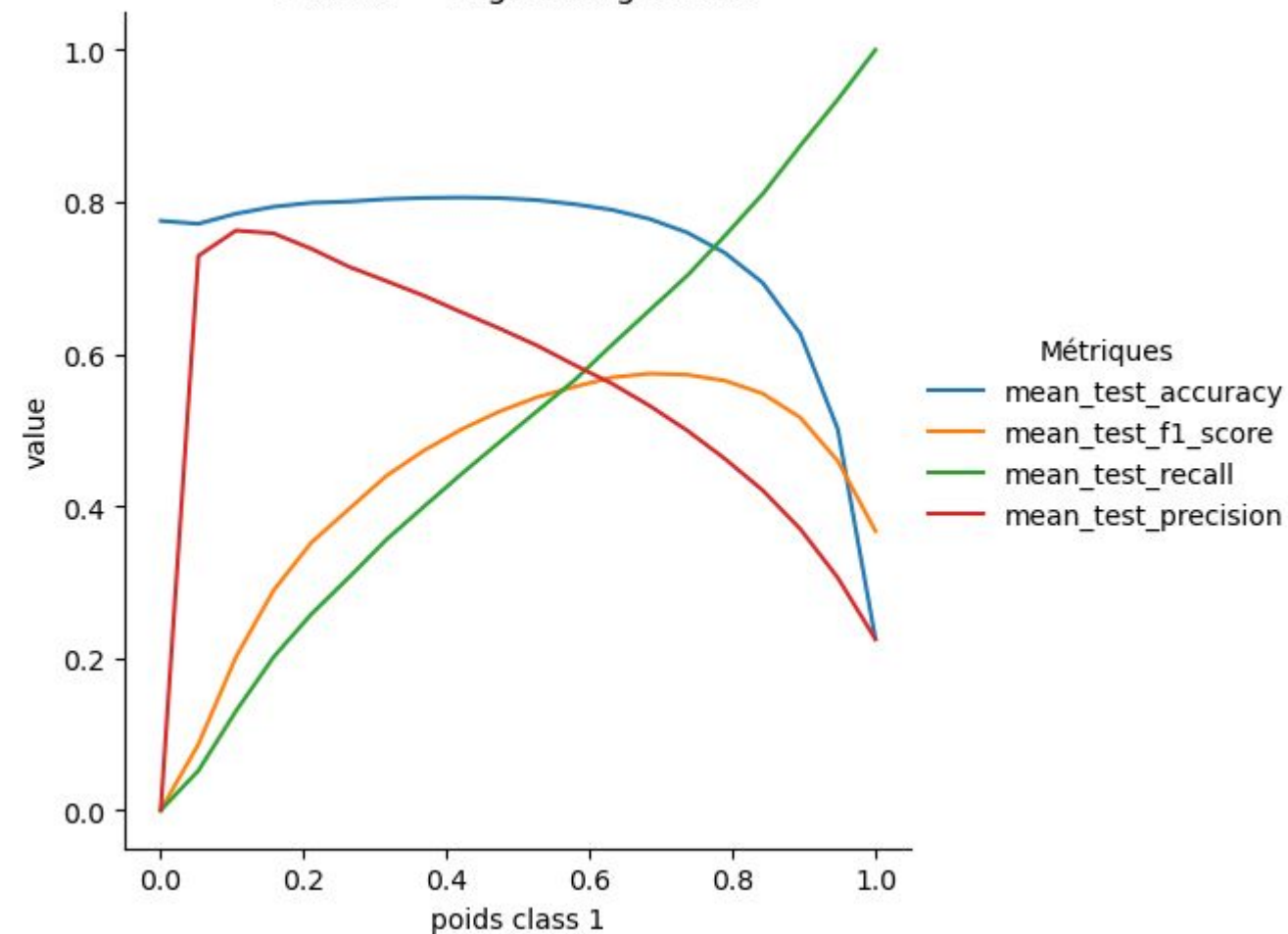
Modèles robustes



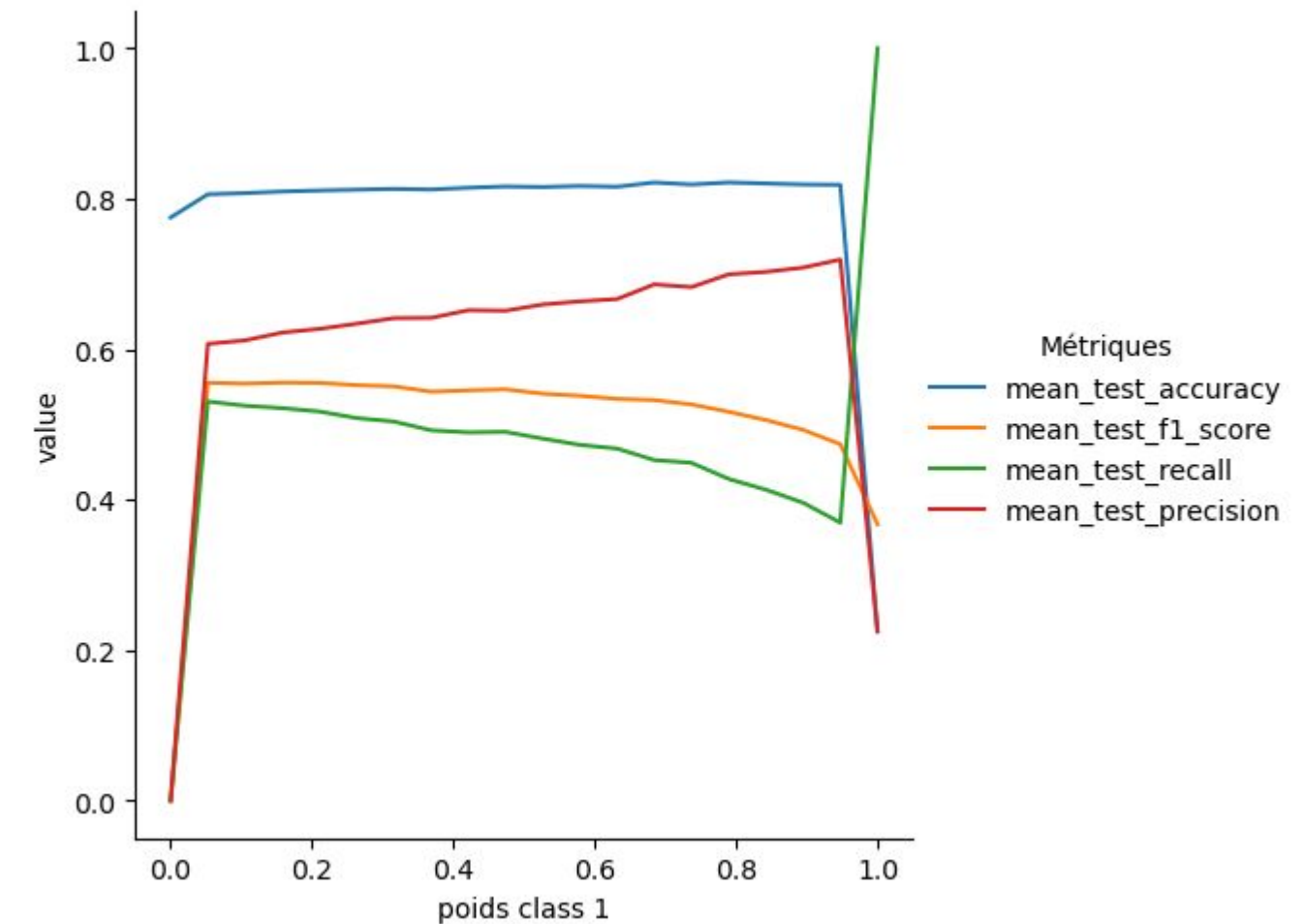
A.3 Annexes : Méthodes de gestion du déséquilibre

Paramètres : class weight

Evolution des métriques de la recherche selon la valeur du poids pour la classe 1
Modèle = LogisticRegression



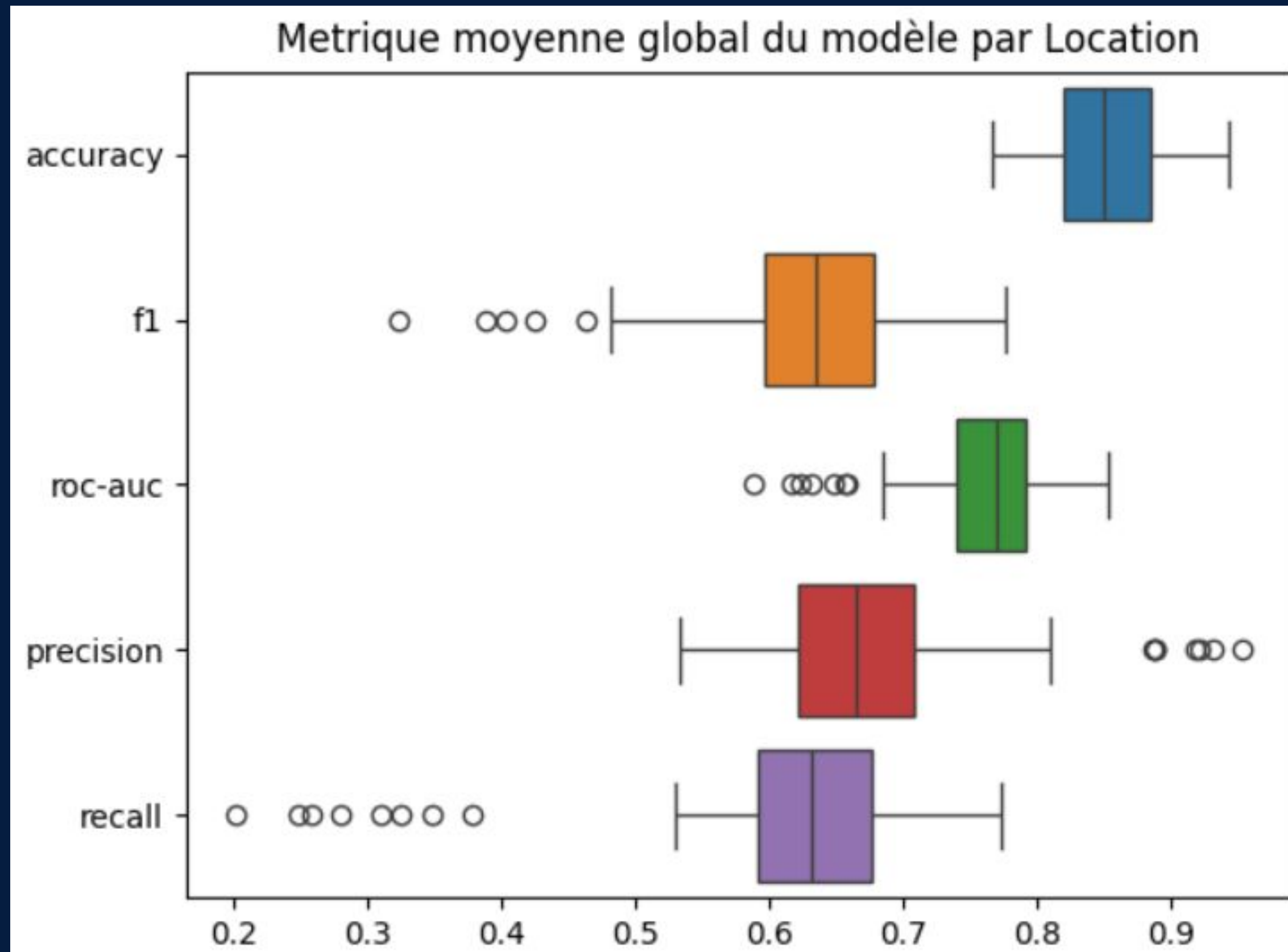
Evolution des métriques de la recherche selon la valeur du poids pour la classe 1
Modèle = RandomForestClassifier



A.4 Annexes : Comparaison des métriques des modèles global et par station

	Modèle global Logistic Regression optimisé sur le f1-score	Modèle global Random Forest optimisé sur le f0.5-score	Modèle global Balanced Random Forest optimisé sur le f1-score	Modèle par station Logistic Regression Moyenne des scores
Accuracy	0.82	0.86	0.81	0.86
F1-score	0.63	0.60	0.64	0.61
précision classe 1	0.58	0.76	0.55	0.69
Rappel classe 1	0.68	0.50	0.76	0.58

A.5 Annexes : Statistiques des métriques par station



3.1 MODÉLISATION DE LA PROBABILITÉ DE PLUIE sur l'ensemble du jeu de données → Streamlit

Métriques de classification adaptées

Classificateurs variés testés

Méthodes de gestion du déséquilibre de classe

- Rééchantillonnage
- Classificateurs robustes au déséquilibre : Balanced Random Forest
- Paramètres : class weight / seuil de classification

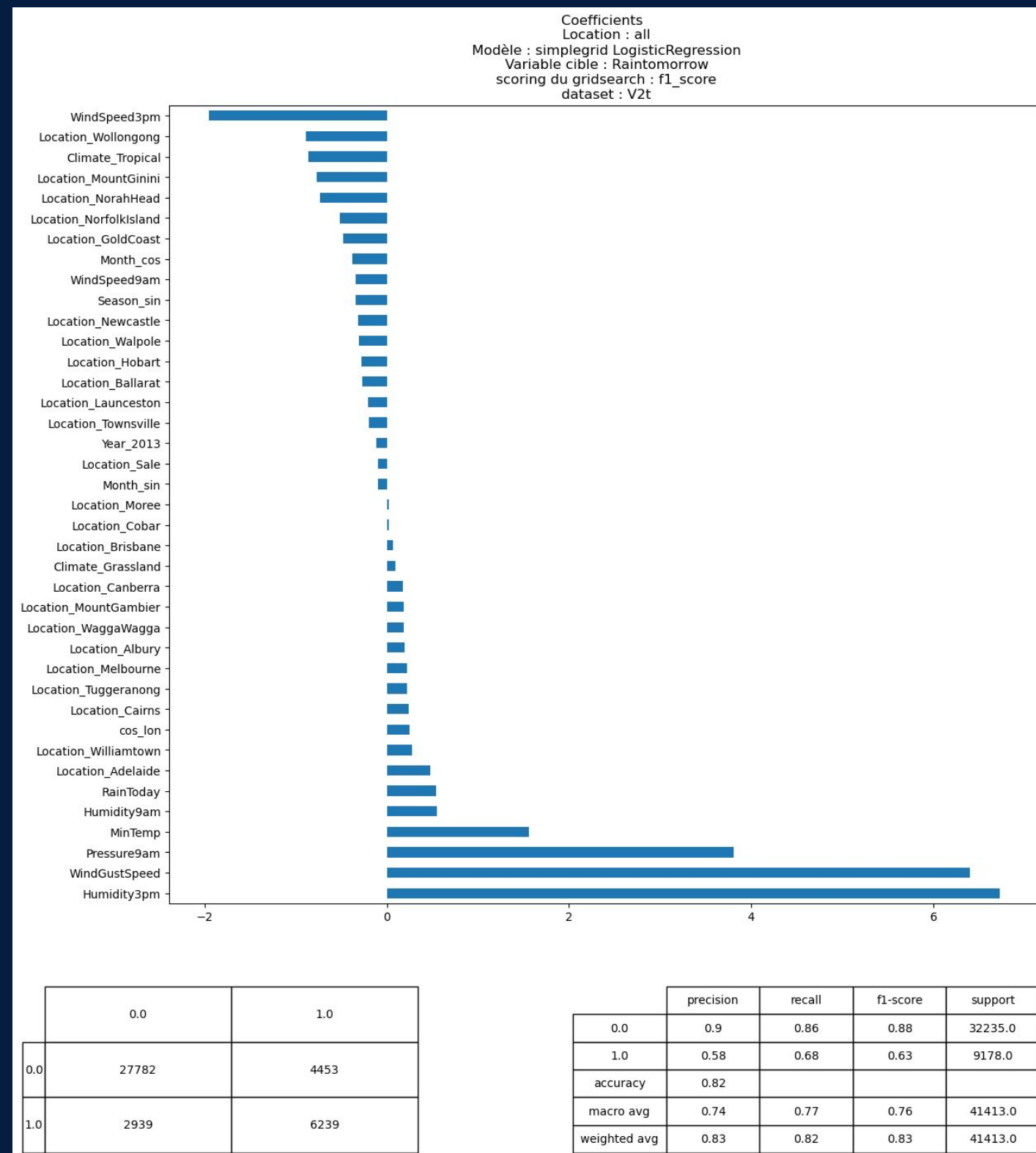
Optimisation selon plusieurs métriques

Tests de plusieurs preprocessing et gestion de la multicolinéarité

3.1 QUELQUES RÉSULTATS → *Streamlit*

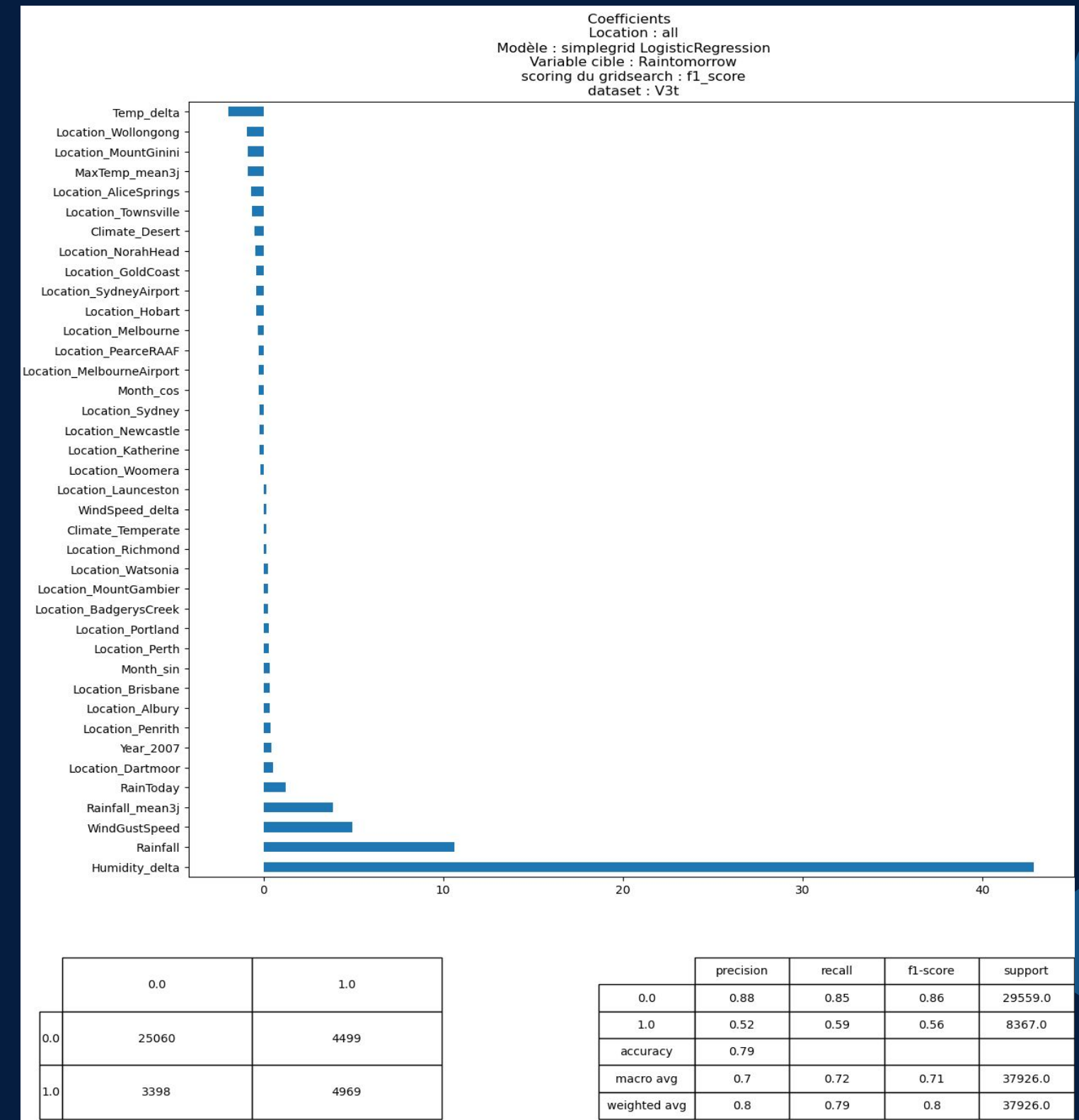
Ensemble des variables

→ Performances



Gestion multicollinéarité

→ Interprétabilité



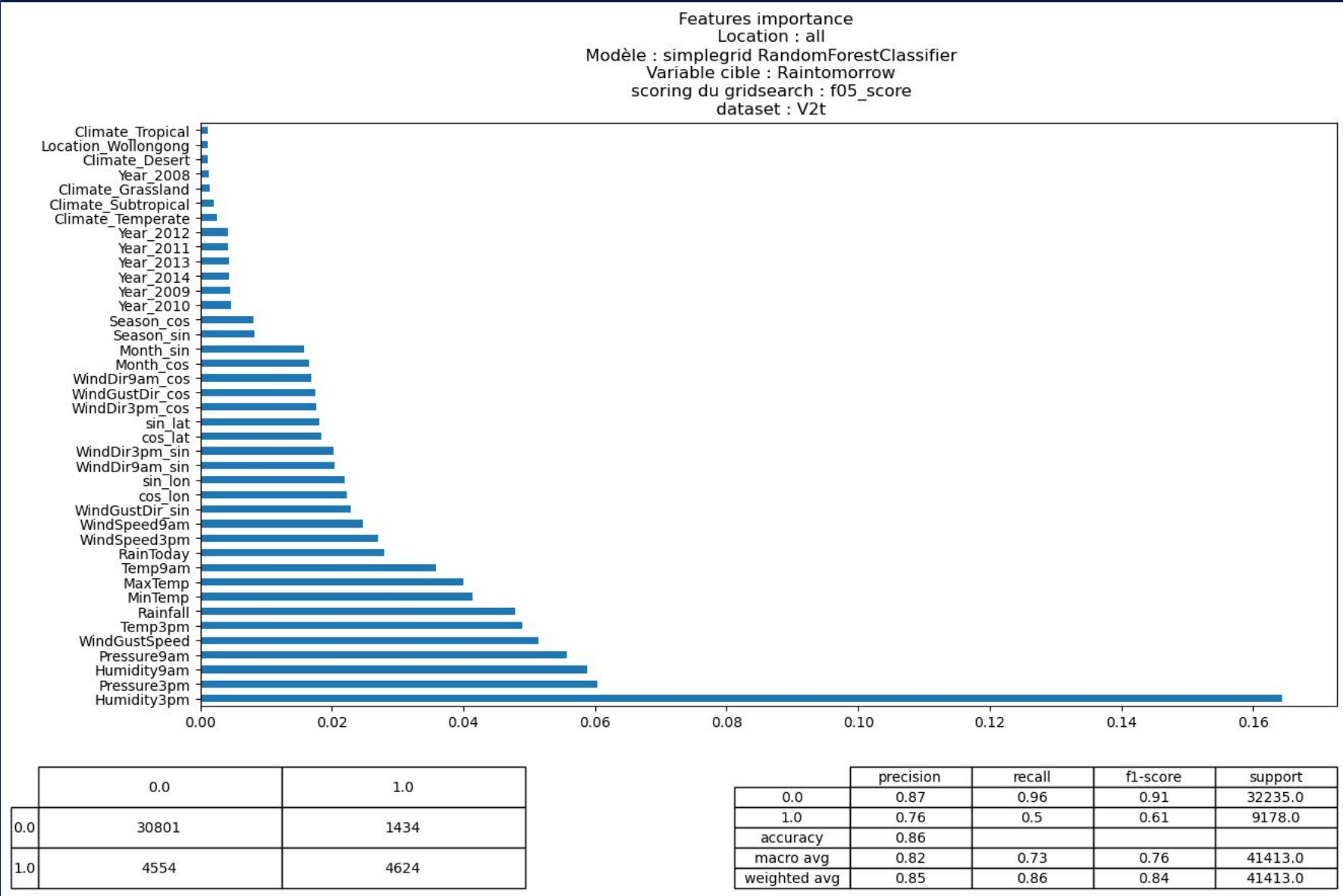
3.1

QUELQUES RÉSULTATS → Streamlit

Random forest

Optimisation f0.5_score

Précision +



Balanced Random forest

Optimisation f1_score

Rappel +

