DÉTECTION DES ANOMALIES

Contrôle de la qualité à l'aide de l'apprentissage machine

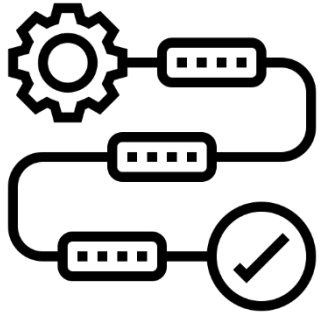
Gurkanwal Arora, scientifique des données
Bureau du dirigeant principal des données (BDPD)

Thinesh Sornalingam, gestionnaire
Service météorologique du Canada (SMC) – Section de la gestion des données et des opérations (SGDO)

Environment and Climate Change Canada
Environnement et Changement climatique Canada
Canada

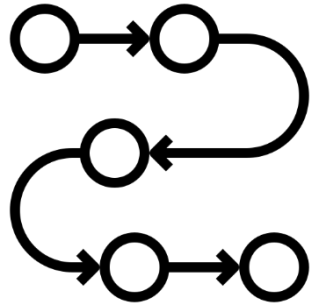
APERÇU DE LA PRÉSENTATION

Problème et objectif



Méthodologie

Résultats

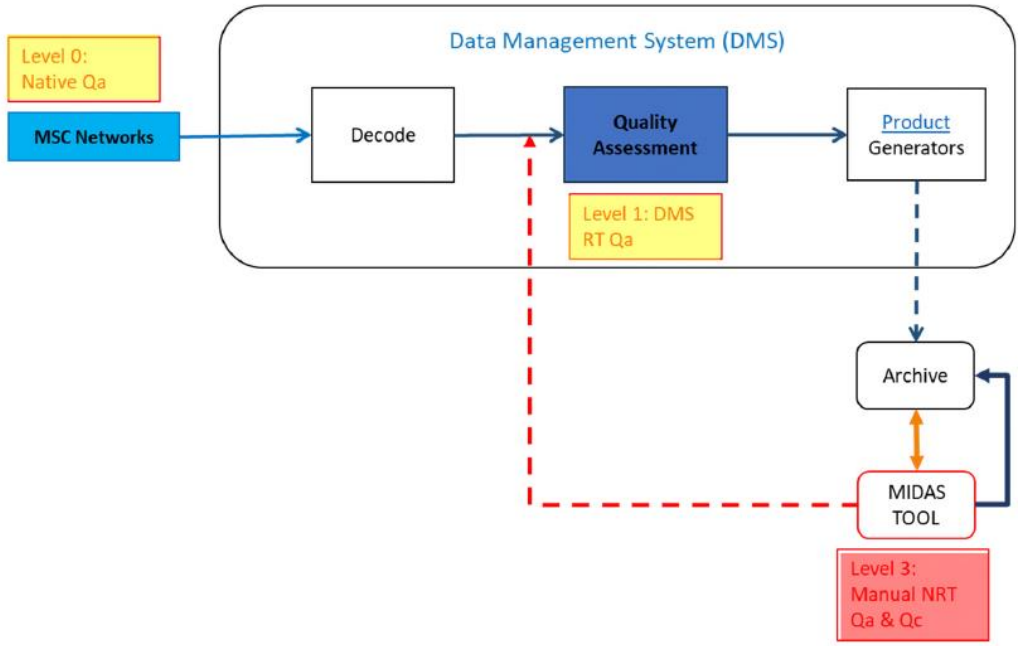


Prochaines étapes

PROCESSUS DE CQ DU SMC

Table 1 An overview of the structure of the DMS QC System.

QC LEVEL	STATUS	MANAGED BY	PROCESSING METHOD	TIMING
Level 0 (i.e. Native)	Operational	DQU (Support from Data Providers)	Automated	Real-Time
Level 1	Operational Further Enhancements in Planning	DQU	Automated	Real-Time
Level 2	Not Started	DQU	Automated	Non-Real-Time
Level 3	Operational	AOU (Support from DQU)	Manual	Non-Real-Time
Level 4	In Development	AOU (Support from DQU)	Auto/Manual	Non-Real-Time



Unité de la qualité des données (UQD) – Wiki (ec.gc.ca)

PROBLÈME ET OBJECTIF

Les observations sont révisées manuellement par les techniciens du contrôle de la qualité (CQ) pour améliorer la qualité des données climatiques.

- Stations et paramètres haute visibilité
- Stations avec des problèmes historiques connus
- Données liées au produit
- Anomalies signalées par les utilisateurs

Des milliers d'observations horaires, et pas assez de ressources humaines pour tout examiner.

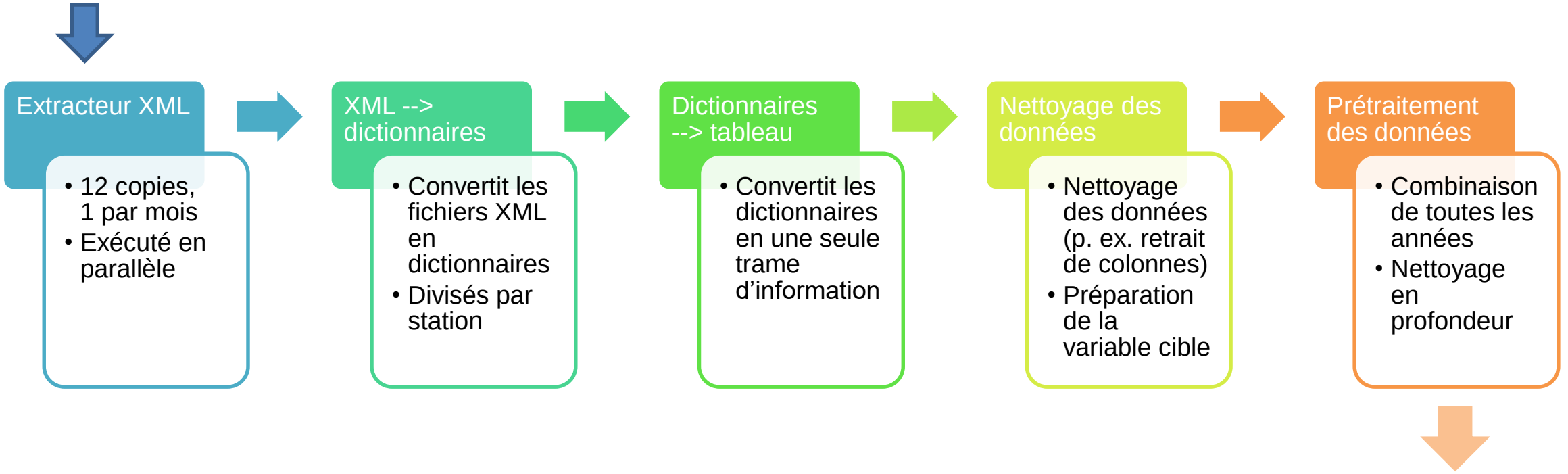
Le SMC cherche une approche plus robuste, automatisée et durable pour détecter les valeurs aberrantes dans les grands volumes de données.

Objectif : Évaluer le potentiel d'utilisation de l'apprentissage machine/de la détection des anomalies pour soutenir le processus de CQ.

* Le modèle d'apprentissage machine contribuera au niveau 3 dans le processus de CQ.

PROCESSUS D'EXTRACTION, DE TRANSFORMATION ET DE CHARGEMENT

Base de données Arkeon



Ensemble de données final

- 1 pour chaque élément
- Inclut toutes les stations et toutes les années
- Prêt pour l'apprentissage machine

ÉTIQUETAGE DES DONNÉES

- L’étiquetage des données fondé sur des règles est intégré dans le traitement des données.
- L’étiquetage est basé sur les corrections manuelles précédentes.

Règles d’étiquetage des données :

1. Soit la valeur de l’indicateur d’assurance de la qualité était « 100 » et a été remplacée par une valeur non nulle, OU la valeur était différente de « 100 » et a été remplacée par « 100 » (cible = 1).
2. La valeur de l’indicateur d’assurance de la qualité après les corrections manuelles est « 100 » en raison d’un remplacement de valeur (cible = 0).
3. Les deux valeurs de l’indicateur sont « 100 », mais ceci est causé par une réexécution du test après un remplacement de valeur (cible = 1).

Table 3 The Data Quality Flags

Flag Value	Flag Name (English/French)	Definition
-10	Suppressed / Réprimer	The data provider has indicated that the value is not to be used or published.
-1	Missing / Manquant	There is no value available.
0	Error / Erreur	The value is erroneous.
10	Doubtful / Douteux	The value may be acceptable but is significantly uncertain.
20	Inconsistent / Incohérent	The value departs significantly from an expected physical relationship with an independently measured, associated variable.
100	Accepted/Passed / Accepté	The value passed all applicable quality assessment test(s) or has been verified as acceptable.

Unité de la qualité des données (UQD) – Wiki (ec.gc.ca)

PRÉTRAITEMENT DES DONNÉES

Valeurs manquantes

- Retrait des colonnes avec plus de 1 % d’instances manquantes
- Utilisation de SimpleImputer (moyenne) de Sklearn pour les valeurs manquantes restantes

Attributs catégoriels

- Encodage factice/encodage 1 parmi *n* (OHE)

MODÉLISATION

Élément	Longueur du champ de données	Nombre de valeurs			Nombre de colonnes	Données de formation (80 %)	Données de test (20 %)
		Classe 0	Classe 1	%			
snow_depth_3022 (épaisseur de neige)	1 718 100	1 705 004	13 096	0,8 %	197	1 374 480	343 620
snow_depth_3025 (épaisseur de neige)	1 702 049	1 688 634	13 415	0,8 %	238	1 361 639	340 410
precipitation_amount_12 (quantité de précipitations)	223 447	199 625	23 822	10,7 %	161	178 757	44 690
wind_speed_3003 (vitesse du vent)	90 085	82 155	7930	8,8 %	337	72 068	18 017

Algorithmes de classification testés :

- Forêt d'arbres décisionnels (RFC)
- Optimisation par gradient (GBC)
- Optimisation adaptative (AdaBoost ou ABC)
- Optimisation par gradient extrême (XGBC)
- Optimisation par gradient léger (LGBC)
- Classifieur de vote (VTC)
- Apprentissage profond

GESTION DU DÉSÉQUILIBRE

- Poids des classes
- Suréchantillonnage
  - Technique de suréchantillonnage des minorités synthétiques (SMOTE)
  - Algorithme synthétique adaptatif (ADASYN)
  - Aléatoire
- Sous-échantillonnage
  - Aléatoire

HYPERPARAMÈTRES

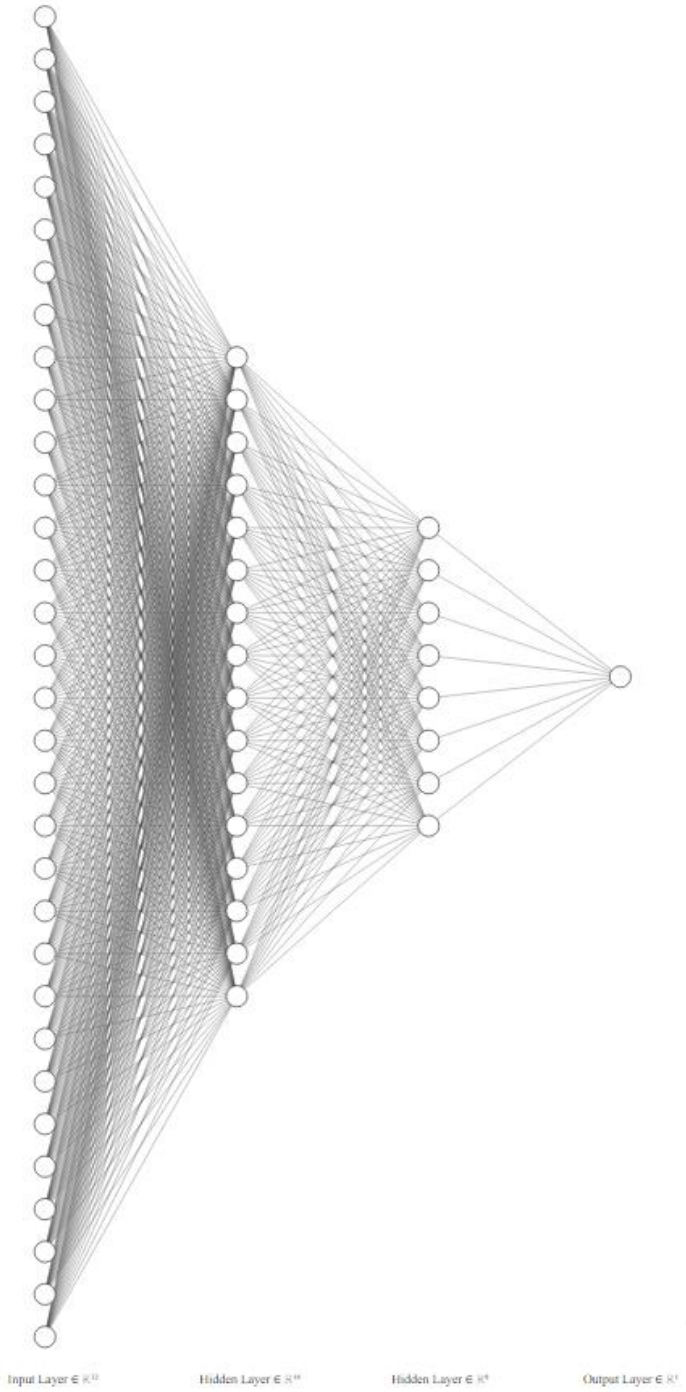
	n_estimators (nombre d'estimateurs)	**max_depth** (profondeur maximale)	**class_weight** (poids des classes)	**learning_rate** (facteur d'apprentissage)	**num_leaves** (nombre de feuilles)
RFC	500	Aucune	Aucun	S. O.	S. O.
GBC	1000	3	S. O.	0,1	S. O.
ABC	1000	S. O.	S. O.	0,1	S. O.
XGBC	1500	Aucune	Aucun	0,1	S. O.
LGBC	500	-1	Aucun	S. O.	31

APPRENTISSAGE PROFOND

Modèle : « séquentiel »

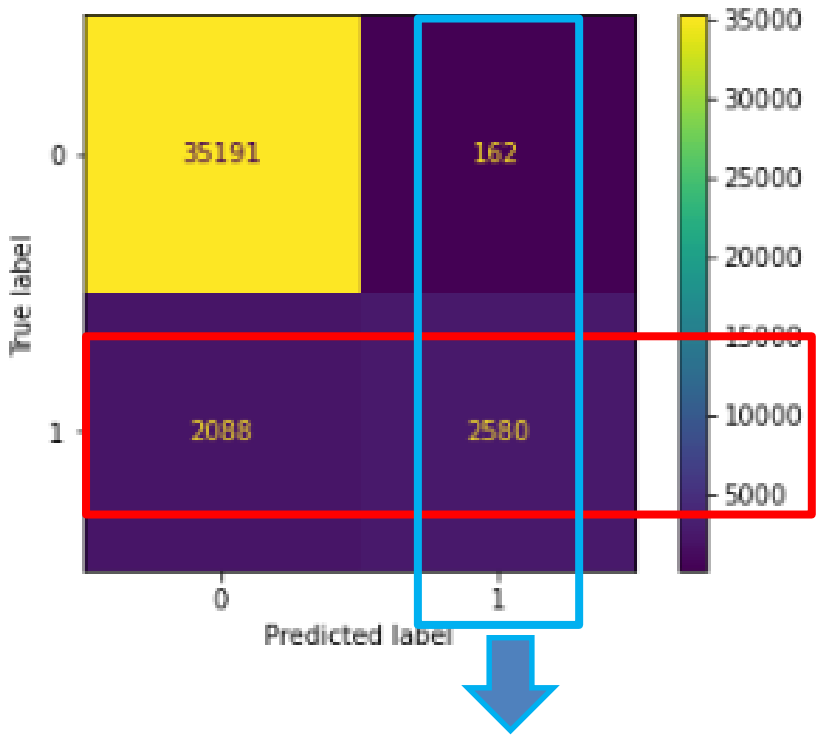
Couche (type)	Forme de sortie	Nb de paramètres	Activation	Perte
dense (dense)	(aucune, 32)	5152	« ReLU »	
dropout (perte)	(aucune, 32)	0		0,2
dense_1 (dense)	(aucune, 16)	528	« ReLU »	
dropout_1 (perte)	(aucune, 16)	0		0,2
dense_2 (dense)	(aucune, 8)	136	« ReLU »	
dropout_2 (perte)	(aucune, 8)	0		0,2
dense_3 (dense)	(aucune, 1)	9	« sigmoïde »	

Nombre total de paramètres : 5825
Paramètres pouvant être entraînés : 5825
Paramètres non entraînables : 0



RÉSULTATS

precipitation_amount_12 – RFC



Classe	0	1	Macro-moy.
Rappel	0,995	0,553	0,774
Précision	0,944	0,941	0,942
Score F1	0,969	0,696	0,833

$$Rappel = \frac{2580}{2580 + 2088} = 0{,}553$$

$$Précision = \frac{2580}{2580 + 162} = 0{,}941$$

Rappel : % des anomalies réelles que le modèle a correctement prédites

Précision : % de toutes les anomalies prévues qui sont correctes

RÉSULTATS

Toutes les stations, données de 2018 à 2021
Scores macro-moy.

Élément	Mesure du rendement	Référence	Apprentissage machine classique				Apprentissage profond
		DummyClassifier*	RFC	XGBC	LGBC	Vote	AP
snow_depth_3022	Rappel	0,499	0,917	0,917	0,918	0,918	0,850
	Précision	0,499	0,995	0,995	0,931	0,931	0,952
	Score F1	0,499	0,953	0,953	0,925	0,925	0,895
snow_depth_3025	Rappel	0,499	0,913	0,951	0,910	0,931	0,817
	Précision	0,499	0,995	0,992	0,947	0,993	0,961
	Score F1	0,499	0,950	0,970	0,928	0,960	0,876
precipitation_amount_12	Rappel	0,502	0,946	0,948	0,951	0,946	0,931
	Précision	0,501	0,959	0,955	0,955	0,959	0,943
	Score F1	0,501	0,952	0,951	0,953	0,952	0,937
wind_speed_3003	Rappel	0,500	0,960	0,973	0,976	0,969	0,909
	Précision	0,500	0,966	0,976	0,979	0,974	0,852
	Score F1	0,500	0,963	0,975	0,978	0,972	0,878

* Le classifieur factice (DummyClassifier) prédit de façon aléatoire en fonction de la distribution.

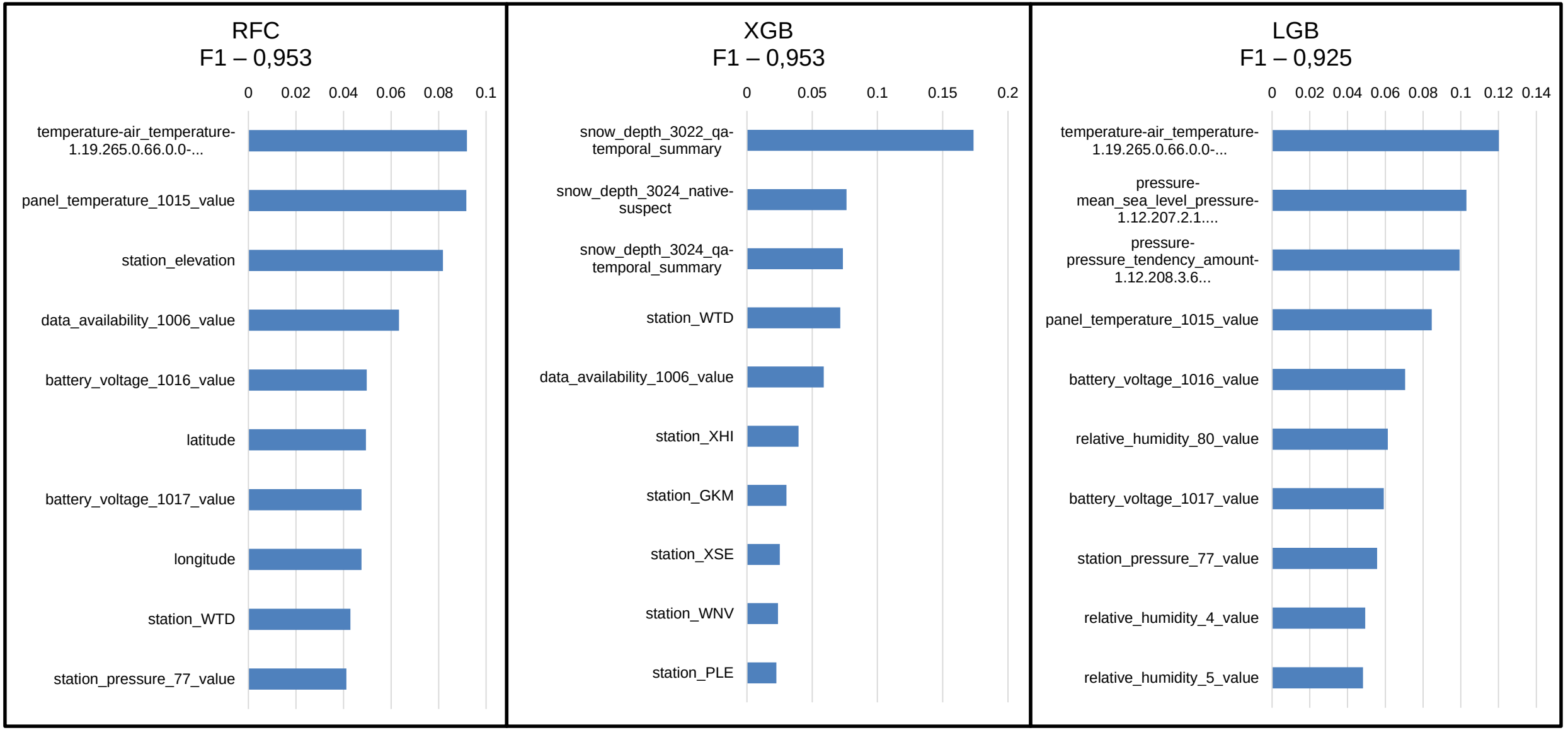
RÉSULTATS

Toutes les stations, données de 2018 à 2021
Scores de la classe 1

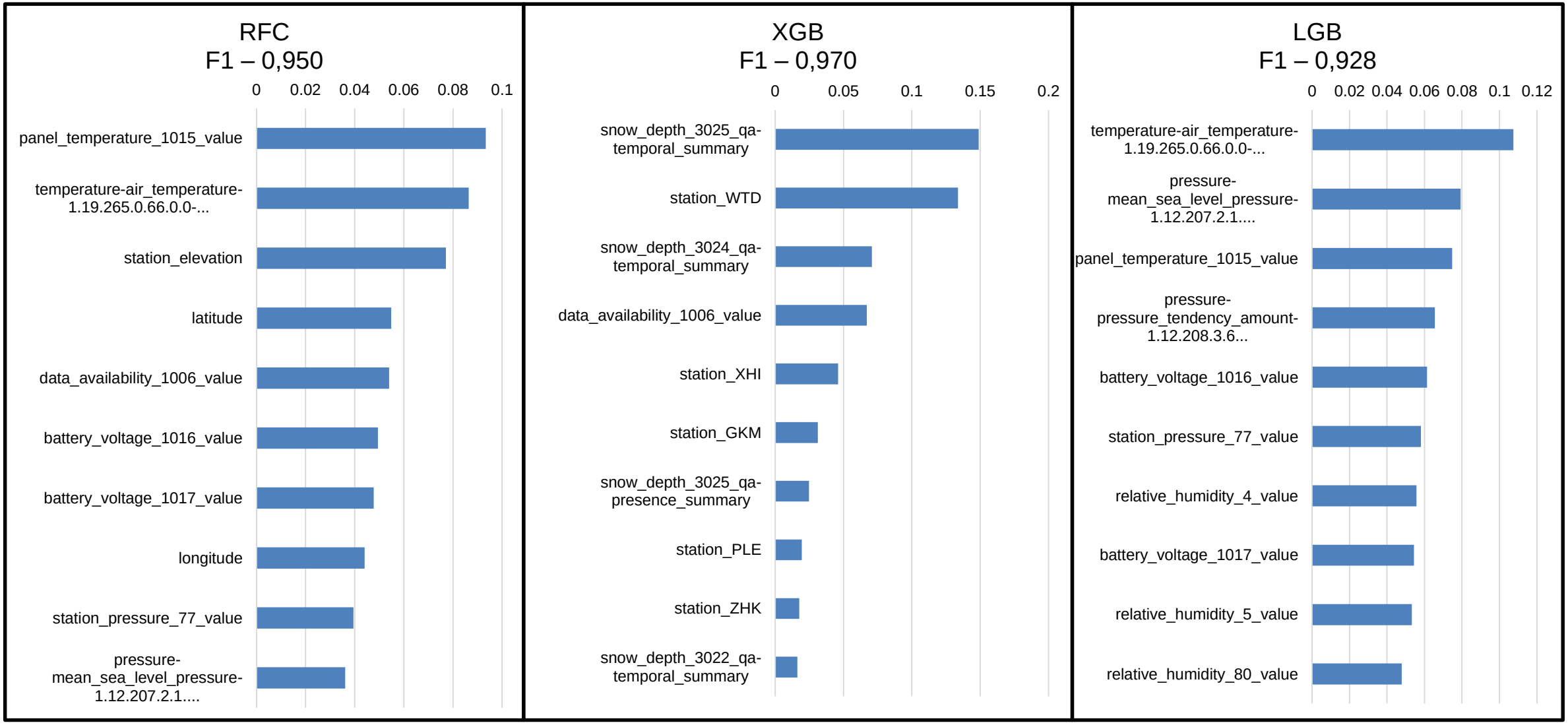
Élément	Mesure du rendement	Référence	Apprentissage machine classique				Apprentissage profond
		DummyClassifier*	RFC	XGBC	LGBC	Vote	AP
snow_depth_3022	Rappel	0,01	0,835	0,835	0,838	0,838	0,701
	Précision	0,01	0,992	0,992	0,864	0,864	0,906
	Score F1	0,01	0,907	0,907	0,851	0,851	0,791
snow_depth_3025	Rappel	0,01	0,825	0,901	0,821	0,861	0,634
	Précision	0,01	0,991	0,984	0,896	0,988	0,926
	Score F1	0,01	0,901	0,941	0,857	0,920	0,753
precipitation_amount_12	Rappel	0,11	0,900	0,905	0,911	0,899	0,872
	Précision	0,11	0,930	0,920	0,921	0,929	0,901
	Score F1	0,11	0,915	0,913	0,916	0,914	0,887
wind_speed_3003	Rappel	0,09	0,925	0,950	0,956	0,943	0,851
	Précision	0,09	0,940	0,958	0,962	0,953	0,718
	Score F1	0,09	0,933	0,954	0,959	0,948	0,779

* Le classifieur factice (DummyClassifier) prédit de façon aléatoire en fonction de la distribution.

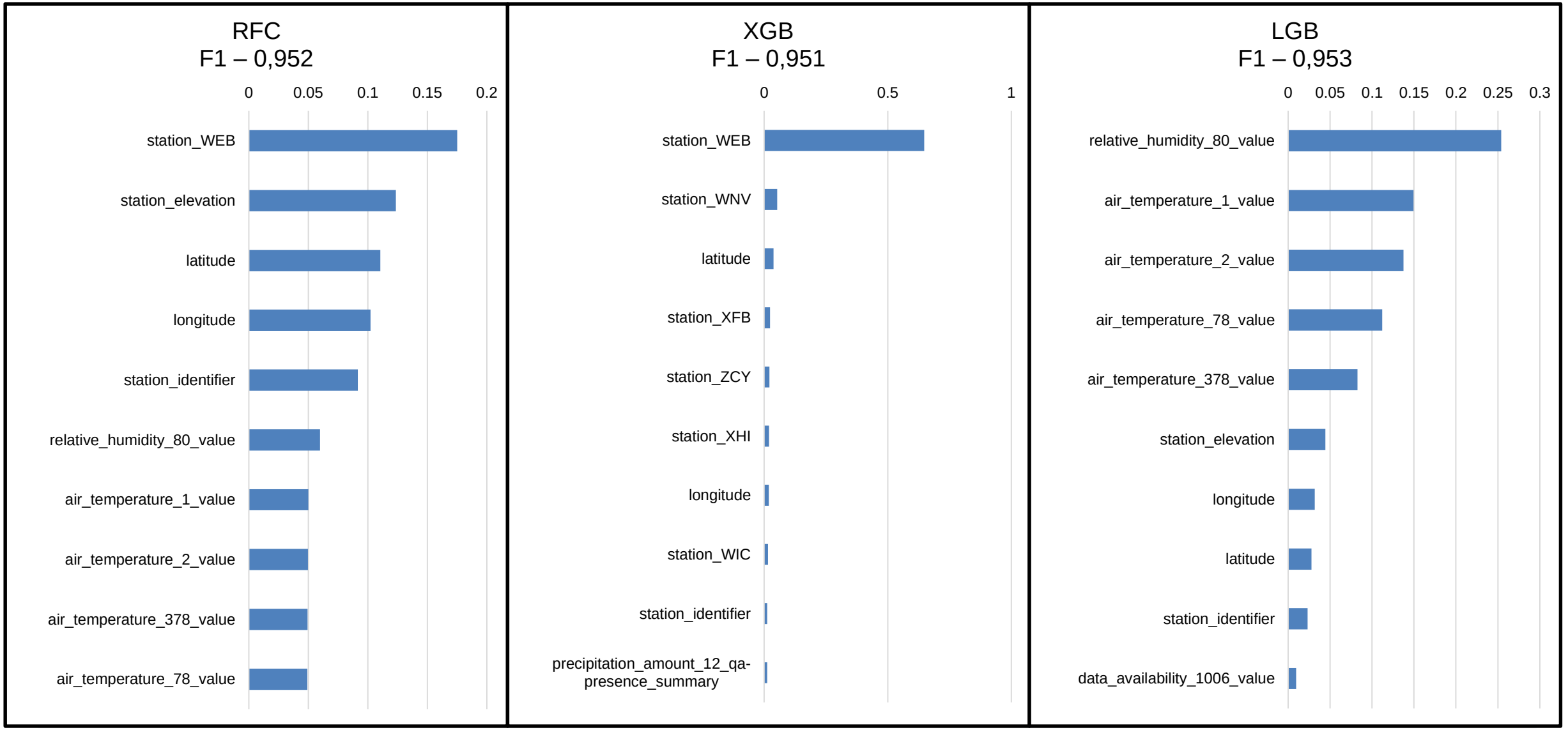
IMPORTANCE DES ATTRIBUTS DE SNOW_DEPTH_3022



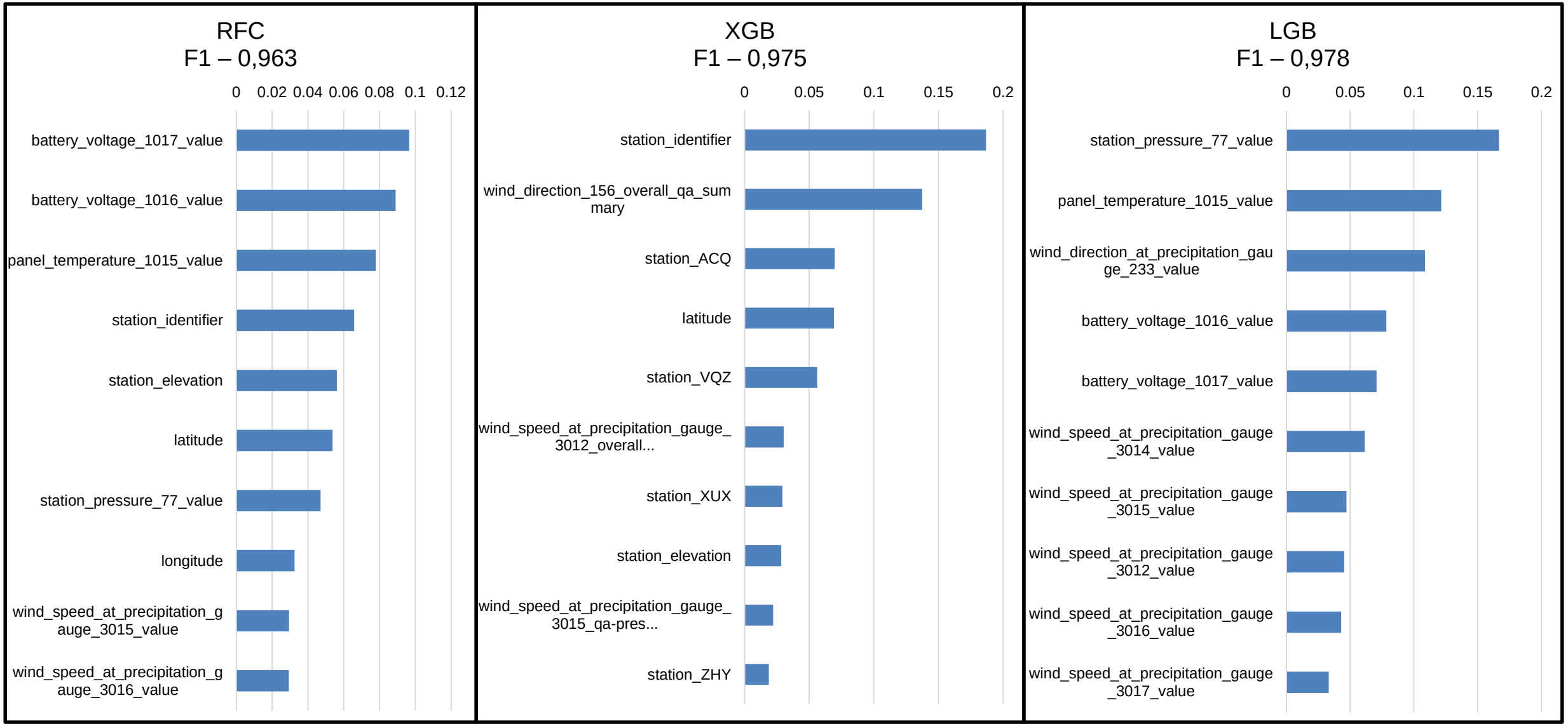
IMPORTANCE DES ATTRIBUTS DE SNOW_DEPTH_3025



IMPORTANCE DES ATTRIBUTS DE PRECIPITATION_AMOUNT_12



IMPORTANCE DES ATTRIBUTS DE WIND_SPEED_3003



PROCHAINES ÉTAPES

- Configurer un environnement pilote pour tester les modèles en production
- Configurer un système de détection et d’alerte en temps réel (alertes par courriel)
- Prévoir des rondes d’essai avec les utilisateurs finaux (techniciens du CQ)