Contrat d'Optimisation du Projet de l'Analyse et Prédiction de la Qualité de l'Air en Inde

Ce contrat d'optimisation du projet vise à améliorer significativement les performances du modèle en augmentant la précision, le F1-score, et la robustesse tout en réduisant la latence d'inférence et la taille du modèle. L'optimisation doit se faire dans le respect de contraintes strictes, incluant un temps d'inférence limité, une taille maximale du modèle, une durée d'entraînement encadrée et l'absence de dégradation des métriques d'équité ou de biais. Le succès sera mesuré principalement via des indicateurs tels que le rappel, la précision, le F1-score ou le RMSE selon le type de modèle, complétés par des métriques secondaires comme le temps d'inférence, le coût d'infrastructure, la stabilité et l'équité. Les ressources allouées comprennent un nombre défini d'heures de travail analyste ou data scientist, l'utilisation d'infrastructures GPU spécifiques et un budget cloud limité. La durée totale prévue pour l'optimisation est encadrée, avec pour livrables un modèle optimisé, un rapport comparatif des métriques avant et après optimisation, ainsi que les scripts d'inférence testés et prêts pour le déploiement, garantissant ainsi une amélioration mesurable et contrôlée des performances du projet.

En résumé, nous avons :

Élément	**Détails**
Projet	Analyse et prédiction de la qualité de l'air en Inde à partir du jeu de données city_day.csv
Objectif principal	Optimiser les modèles de classification pour améliorer la performance prédictive de l'AQI_Bucket_Encoded : - Augmenter le F1-score de 0.8312 à ≥ 0.85 - Réduire le Log-Loss de 0.4472 à ≤ 0.42 - Stabiliser les performances sur les villes avec forte variabilité de pollution
Contraintes	- Temps d'entraînement maximal par modèle : 4 heures - Taille du modèle $\leq$ 200 Mo - Pas de dégradation de la précision pour les

	catégories moins représentées d'AQI - Temps d'inférence ≤ 1 seconde par observation
Métriques primaires	- F1-score - Log-Loss - Accuracy globale
Métriques secondaires	- Temps d'inférence - Stabilité du modèle (variance des métriques sur k-folds) - Importance et interprétabilité des features - Biais géographique ou saisonnier
Budget de calcul et de travail	- 1 GPU RTX 4080 disponible pendant 3 jours - 20 heures d'expérimentation - Budget cloud maximal : 300 $
Durée estimée	2 semaines pour expérimentation et optimisation complète
Modèles concernés	XGBoost (priorité) - Random Forest - K-Nearest Neighbors (KNN)

A noter :

1. Le **modèle XGBoost optimisé** reste la priorité, mais RF et KNN seront évalués après ajustement pour garantir la robustesse.
2. L'**optimisation inclura** :
    - Ajustement des hyperparamètres via GridSearchCV ou Optuna.
    - Feature engineering avancé (interaction entre polluants, variables temporelles et géographiques).
    - Évaluation continue sur l'ensemble de test et sous-ensembles représentatifs (par villes et saisons).
3. **Livrables attendus** :
    - Modèle final XGBoost optimisé avec métriques validées.
    - Tableau comparatif des modèles avant et après optimisation.
    - Rapport technique détaillant les hyperparamètres choisis, les performances et les éventuelles limites.