Université Hassan II de Casablanca
Ecole Nationale Supérieure d'Arts et Métiers de Casablanca

Nom et Prénom: Mohammed AMEKSA
Filière: Master Big Data et Internet des Objets

Sujet: "Benchmark de l'impact de la plateforme et la technique Machine Learnin sur la performance du modèle développé: Cas d'un problème de Régression"

Résumé

La visibilité réduite représente un défi pour les prévisionnistes météorologiques depuis longtemps, vu son impact néfaste sur la circulation aérienne, maritime et routière. En effet, les pertes humaines et financières attribuables à des baisses de visibilité sont devenues de plus en plus accentuées, d'où une bonne prévision de la visibilité horizontale a un grand apport pour les prévisionnistes météorologiques. Ainsi, pour faire face à ce déficit dans le domaine de la prévision numérique du temps, le potentiel des techniques du Datamining à estimer la visibilité horizontale a été évalué dans plusieurs études scientifique, cependant, les performances des modèles développés diffèrent d'une étude à une autre due à la diversité des outils et des algorithmes du Datamining utilisés. De ce fait, l'objectif de notre étude est d'étudier la sensibilité de la performance des modèles développés à la plateforme et à l'algorithme du Datamining pour un cas de régression visant à estimer la visibilité à partir des prévisions d'un modèle numérique opérationnel AROME. Pour atteindre cet objectif, nous avons utilisé deux familles d'algorithmes, ceux qui se base sur les méthodes ensemblistes y compris adaptatives (gradient boosting, eXtreme gradient boosting) ou aléatoires (random forest) et d'autre qui se base sur l'apprentissage profond (deep learning). Ces techniques Datamining sont évalué sous diverses plateformes open source (Scikit-learn, H2O, WEKA, Tensorflow et Keras). En outre, une base de données qui couvre les données horaires de 3 ans, résultat d'un prétraitement des sorties brutes du modèle de prévision numérique AROME et des données observées a été utilisée dans ce travail.

L'échantillonnage de ces données en 70% d'apprentissage et 30% de test a été effectué en garantissant la représentativité des mois, des heures et des diverses classes de visibilités pour toutes les stations synoptiques. Les résultats de l'évaluation de la sensibilité des modèles développés à la plateforme et l'algorithme utilisé montre que la performance des modèles ensemblistes est la meilleure quelque soit la plateforme utilisée sauf pour Keras où seul le Deep Learning a été utilisé. D'autre part, l'algorithme Random Forest s'affiche comme meilleur estimateur de la visibilité après réglage des hyperparamètres pour les plateformes WEKA et Scikit-learn. Cependant, Gradient boosting machine s'est distingué comme meilleur estimateur pour la plateforme H2O. Les ordres de grandeurs des erreurs enregistrées sont similaires entre les diverses plateformes. Ainsi, on constate des erreurs quadratiques moyennes de 1933 m, 1942 m et 1945 m respectivement pour Gradient Boosting sous H2O, Random Forest pour Scikit-learn et WEKA. De même pour l'erreur absolue moyenne qui prend les valeurs suivantes 1199 m, 1221 m et 1232 m pour les mêmes algorithmes et plateformes.

Mots clés: Régression, Visibilité réduite, Random Forest, Gradient Boosting Machine, eXtrem Gradien Boosting, réseau de neurones, grid search, random search