**Comparaison des classements des villes par type de climat**

Une classification des différentes villes d'Australie a été réalisé pour les séparer selon les types de climat. LLes informations sur le type de climat ont été incorporées dans la base de données et ses performances ont été analysées sur la base de 3 modèles d'apprentissage automatique :

* -un modèle global, un algorithme unique pour toutes les villes et sans intégrer d'information sur le type de climat.
* -un modèle global avec Cluster, c'est-à-dire un modèle global mais en ajoutant une colonne indiquant le type de climat.
* -Un modèle pour chaque cluster. Un modèle pour toutes les villes du même cluster.

Pour les 3 cas, KGBOOST et Randon FOrest ont été utilisées.

Les résultats des différentes métriques montrent un résultat très similaire entre les 3 méthodes. Il semble que l'incorporation des informations sur les clusters améliore légèrement la courbe ROC avec les deux algorithmes, mais cela n'est pas vérifié avec les autres mesures telles que le score ou F1.

Résultats moyens avec XGBOOST :

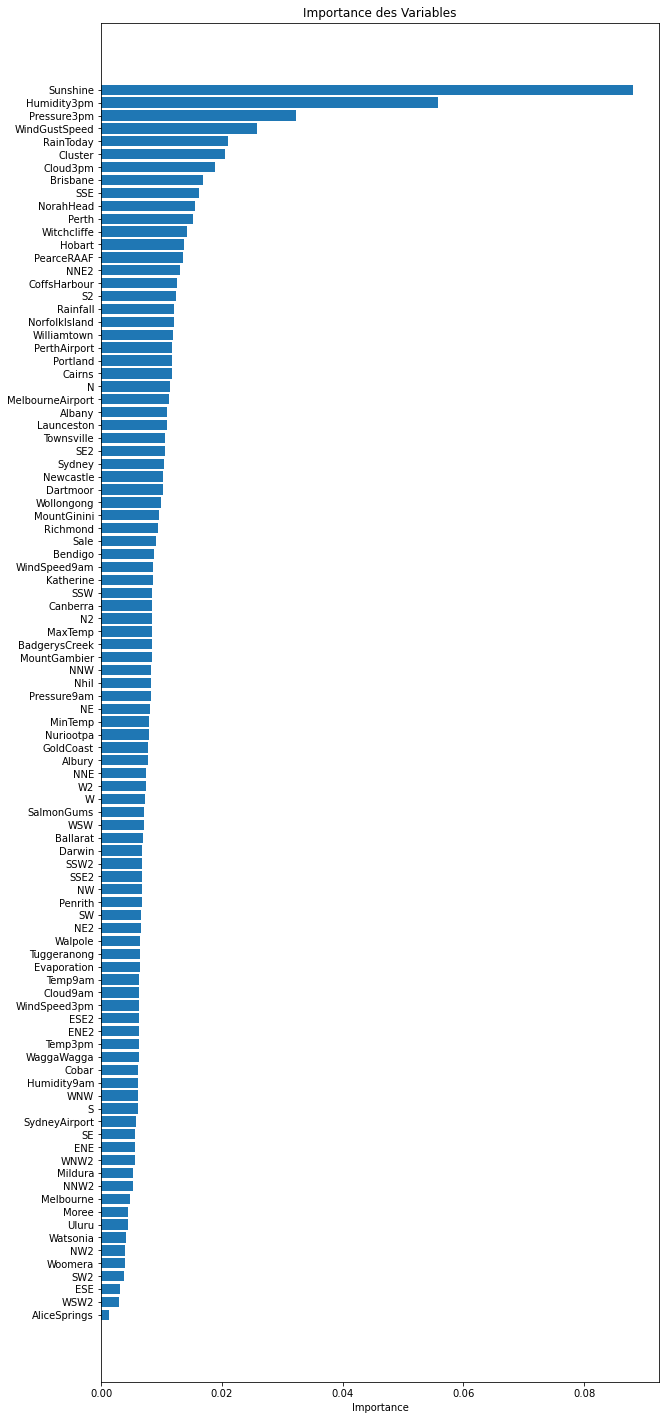
|  | Global | Global avec Cluster | Modèle par Cluster |
| --- | --- | --- | --- |
| Score | **0.8842** | 0.8841 | 0.8405 |
| F1 | **0.7032** | 0.7017 | 0.7002 |
| ROC | 0.9245 | 0.9250 | **0.9259** |

Résultats moyens avec Random Forest :

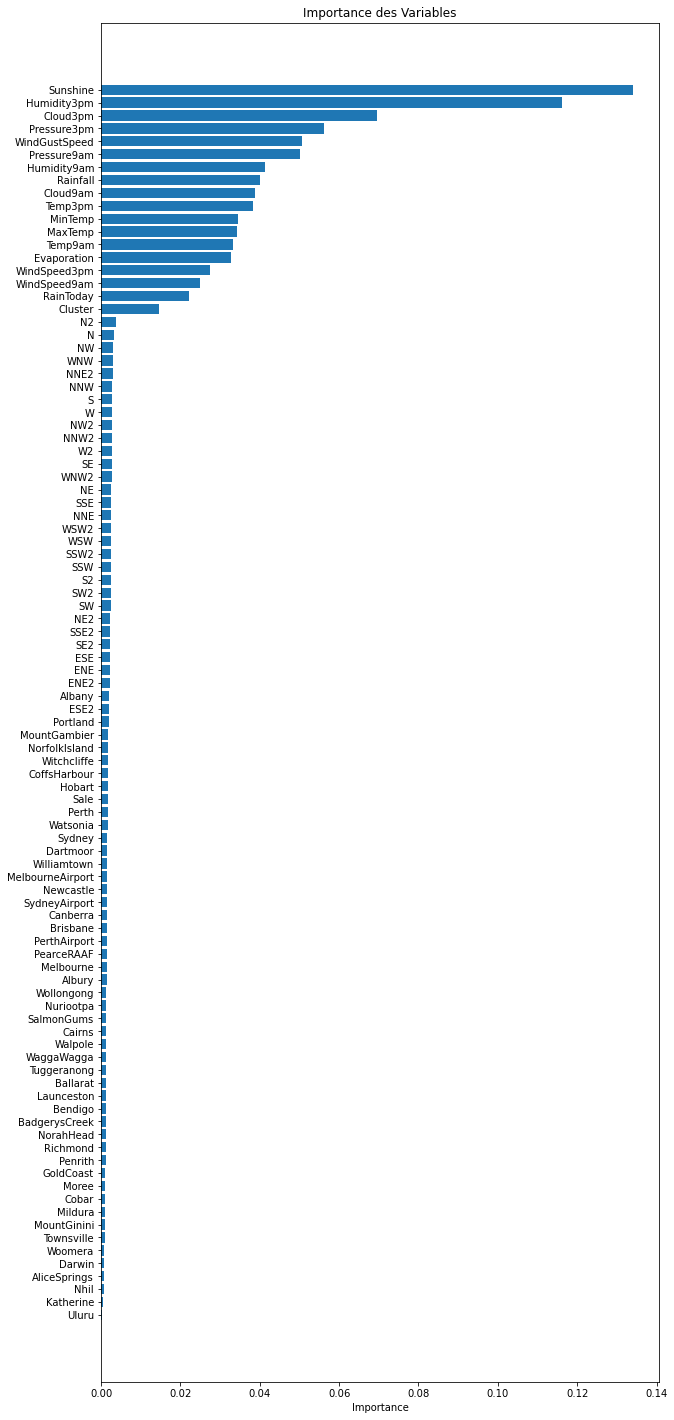
|  | Global | Global avec Cluster | Modèle par Cluster |
| --- | --- | --- | --- |
| Score | 0.8775 | **0.8784** | 0.8454 |
| F1 | 0.6632 | 0.6655 | **0.6668** |
| ROC | 0.9167 | 0.9182 | **0.9191** |

Lorsque l'on regarde l'influence de chaque variable sur la prédiction, on constate que la colonne "Cluster" ne fait pas partie des 4 variables les plus influentes, et que son influence diminue sensiblement par rapport à la variable "sunchine", ce qui peut expliquer que l'amélioration ne soit pas très significative. Sans que cela se traduise nécessairement par une meilleure prédiction, on observe qu'avec l'algorithme GBOOST la variable cluster a plus d'influence que dans la forêt de randon. Obtenant la sixième position en importance, c'est une variable intéressante à inclure pour renforcer le modèle.

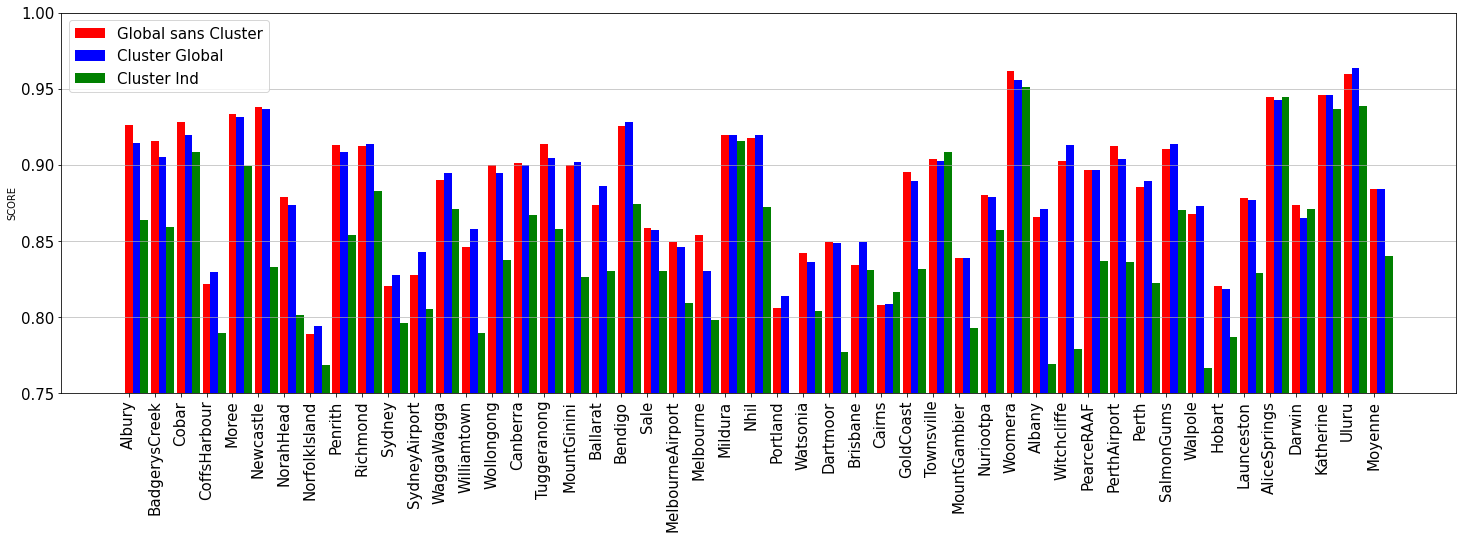
Importance variable XGboost:



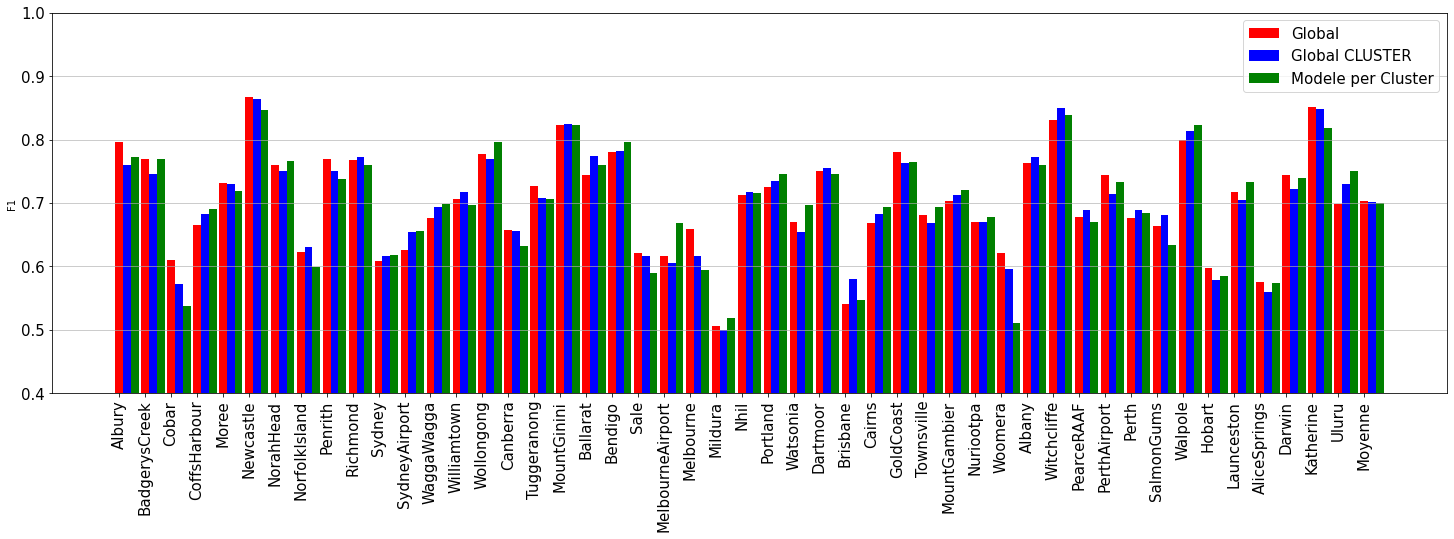
Importance variable Randon Forest:



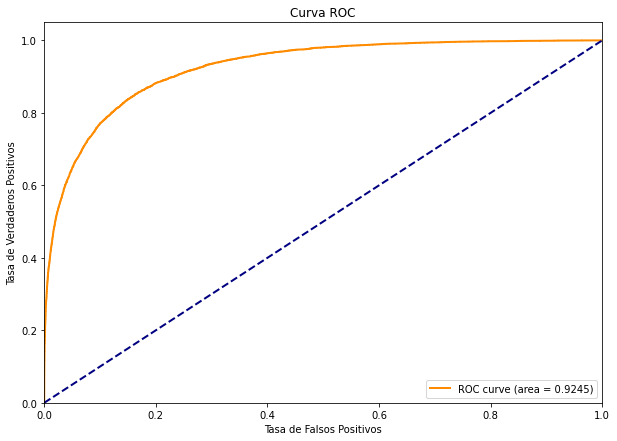
Graph Score par ville.



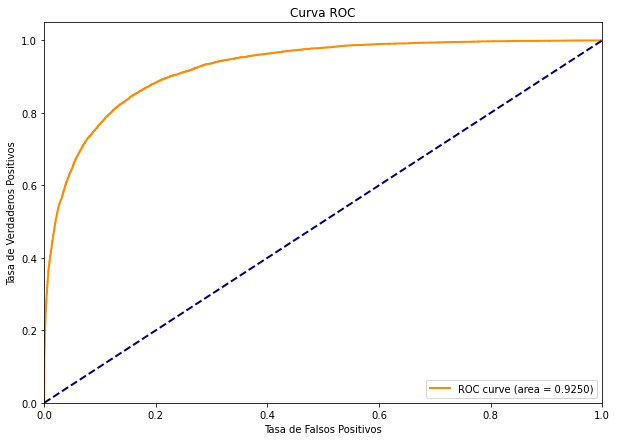
Graph F1 par ville



Curve ROC Modele global sans cluster:



Curve ROC Modele global sans cluster:



Curve ROC Modele global sans cluster:

