

## TP 3 – Forêt Aléatoire

### Librairies nécessaires

ggplot2, rpart, randomForest, plotly, corrplot, pdp

### Données : [MedicalPremium.csv](#)

Analyser rapidement la base de données.

### Compréhension du modèle

1. Quels sont les hyperparamètres à renseigner dans l’algorithme randomForest ?
2. Quelles sont les outputs de la fonction ?
3. Entrainer une première forêt sur la base pour prédire la prime en fonction du reste des variables.

### Interprétation du modèle

1. Visualiser la feature importance à l’aide de varImpPlot. Qu’en déduire ?
2. Analyser l’erreur OOB de la forêt en appliquant plot au modèle entrainé.
3. Evaluer l’impact des différentes variables sur la prédiction grâce au graphique de dépendance partielle (partialPlot).
4. Afficher un arbre de la forêt en particulier grâce à getTree

### Tuning des hyperparamètres

1. Quels hyperparamètres sont à optimiser ? Quelle est la mesure d’erreur à analyser ?
2. Optimiser les deux hyperparamètres un par un en laissant l’autre fixe.
3. S’assurer d’avoir trouvé le couple optimal par grid search.

### Comparaison CART et RF

1. Séparer la base initiale en 10 folds afin d’analyser le pouvoir prédictif des deux modèles par cross-validation. Quelle mesure utiliser ?
2. Estimer le pouvoir prédictif du CART. Quelle est la profondeur de l’arbre ? Comparer par rapport aux arbres de la forêt aléatoire.
3. Estimer le pouvoir prédictif de la forêt aléatoire. Comparer avec le CART.