Ouissame Tahani

Projet noté

Introduction :

Big Datext, entreprise Grenobloise spécialisée dans l'analyse prédictive, et la mairie de Grenoble se sont associées pour la mise en place et la diffusion d'une base de données pour un défi associé à une conférence nationale (EGC 2017). Big Datext et les services de la Ville ont axé le défi sur les données relatives aux espaces verts.

Objectif :

Vous avez à disposition un jeu de données sur les arbres de la ville de Grenoble. Plusieurs mesures ont été relevées sur les arbres. En reposant sur ces données, vous allez prédire si les arbres sont sains ou atteint d'un défaut. Vous pouvez par la suite identifier la localisation des défauts (racines, tronc…).

- Premier but :

  Identifier les arbres ayant un défaut et prédire la localisation de celui-ci

- Deuxième but :

  Connaître l'état du parc végétal afin de mieux comprendre son évolution et fournir des préconisations pour faciliter son entretien.

Appliquons la fouille de données au service du développement durable !

Consignes :

Le but de ce projet est d'acquérir une expérience enfouille de données incluant la préparation des données, le choix et l'évaluation comparative de plusieurs méthodes de fouille. Le projet consiste principalement à :

a. Explorer les données pour déterminer quelles méthodes et quels descripteurs (attributs) sont susceptibles d'être intéressants pour le problème de fouille posé. Vous pourrez utiliser des outils d'analyse exploratoire de données pour produire des graphiques et des statistiques résumant les données.

b. Choisir quelques méthodes à tester sur la base des résultats de votre analyse exploratoire. Une méthode peut être un algorithme précis (dans un logiciel précis) ou une façon de représenter les données ou de les préparer (façon de construire les attributs, de sélectionner des attributs…) ou une combinaison des deux. Vous avez le choix d'utiliser les algorithmes que vous souhaitez :

arbres de décision, règles de classification, méthodes d'ensemble pour la classification (bagging, boosting, forêts aléatoires), clustering (hiérarchique ou à base de densité), extraction de diverses catégories de motifs ensemblistes ou séquentiels.

c. Evaluer de façon rigoureuse les résultats obtenus et donner une interprétation de ces résultats. Si plusieurs méthodes ont été testées (partie prédictive), quelle méthode semble être la meilleure pour le problème posé ? Vous pouvez aussi avoir des idées de choses à faire si vous aviez plus de temps et les mentionner dans le rapport.

Le projet sera noté sur la base d'une présentation orale ou chacun des membres du groupe devra s'exprimer et répondre à quelques questions et d'un rapport final qui devra être rédigé avec soin. Ce rapport devra comporter au moins ces sections :

- Introduction et présentation du projet
- Préparation des données
- Analyse exploratoire des données (cette section peut être fusionnée avec la précédente)
- Programmes de fouille utilisés et paramètres choisis (rappeler les objectifs et principes de ces programmes) en précisant la librairie utilisée (Python, R, Weka, Knime…)
- Résultats, évaluation, comparaison
- Conclusion et discussion

Critères d'évaluation :

Pertinence des méthodes choisies, qualité des expériences effectuées, rigueur et qualité de l'évaluation/comparaison des différents algorithmes, quantité de travail effectué, clarté des explications fournies.