Université de Lorraine - TN 3A IAMD – M2 IMSD - Module **FDEC** - 2019/2020

**Projet FDEC**

# La fouille de données au service du développement durable

*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. Le but du défi est double.

**Défi 1 :**
Il consiste en deux tâches de prédiction visant à déterminer, à partir des données disponibles, si l'arbre a ou non un **défaut** et dans l'affirmative lequel, sachant qu'un arbre peut présenter un défaut à différents endroits : **racine, tronc, collet, houppier.**

**Tâche supervisée 1 : classification uni-label**
Pour prédire au mieux qu'un arbre a un défaut. C'est un problème de classification uni-label car chaque arbre a un seul label défaut.

**Tâche supervisée 2 : classification multi-label**
Pour prédire au mieux les localisations des défauts d'un arbre. Il s'agit d'un problème de classification multi-label puisqu'un arbre peut avoir le défaut au niveau de la racine et du tronc par exemple. Une possibilité est ici est de **construire autant de classifieurs que de classes** (un classifieur pour prédire qu'un arbre a un défaut au collet ou non, un autre classifieur pour prédire qu'un arbre a un défaut à la racine ou non etc.)

Pour information, sur la tâche de prédiction uni-label un classifieur *baseline* permet d'obtenir 86% pour l'exactitude (*accuracy*), 82% de précision et 72% de rappel tandis que sur la tâche multi-label les taux sont respectivement de 64% et 37% pour la précision et le rappel[1].

**Défi 2:**
La seconde tâche, plus ouverte, vise à mieux connaitre l'état du « parc végétal » de Grenoble, mieux comprendre son évolution et fournir des préconisations pour faciliter son entretien. Pour cette seconde tâche, il est possible d'avoir recours à des données externes, de proposer des possibilités de visualisation etc.

**Les données :**
Les données concernent des arbres situés dans la ville de Grenoble et entretenus par les services municipaux. Pour chaque arbre, on dispose de variables décrivant son type, son stade de développement, sa localisation et son environnement, son état et les traitements préconisés à l'occasion de diagnostics.

Le jeu de données est déposé sur Arche. Un deuxième document contient la description des attributs du jeu de données.

[1] La précision est calculée comme la moyenne des précisions des classifieurs dédiés à chaque classe (Tronc, Houppier etc.). Idem pour le rappel.

**Déroulement du projet FDEC :**

Le projet se fera par binôme ou trinôme. Chaque membre du groupe devra justifier d'au moins 15 heures de travail. Chaque binôme devra aborder les deux parties du défi.

Un rapport devrait être rédigé sur le travail effectué avec au moins la description :

- de l'exploration et la préparation des données,
- des algorithmes appliqués aux données, une description plus complète est attendue si l'algorithme n'a pas été vu en cours de FDEC
- la justification du choix de ces algorithmes
- les évaluations comparatives rigoureuses ainsi que l'interprétation qualitative des résultats. Attention notamment à ne pas ignorer les classes minoritaires (peu représentées) lors de l'évaluation.

Ne pas hésiter à mettre en avant d'autres aspects liés au domaine d'application par exemple en quoi et comment vos résultats (tout ou partie de) peuvent être utiles aux botanistes ou aux décideurs de la ville.

Le rapport devra être déposé sur Arche. Une soutenance orale de 10 minutes sera programmée. Vous serez évalué(e)s sur la quantité et le sérieux du travail fourni, la clarté de vos explications (par écrit et en réponse aux questions à l'oral) et la qualité de votre démarche.

**Calendrier et modalités**

Les groupes doivent être faits pour **vendredi** 13/12/2019 à 12h

Envoyer la composition du groupe à Malika.Smail@loria.fr avant cette date

Les rapports devront être déposés sur arche avant **lundi 10/12/2019 à 12h** (avec comme sujet "[Rapport projet FDEC]"

Les **soutenances** seront programmées la **semaine du 8 février**. L'ordre de passage vous sera communiqué avant.

Votre enseignante est accessible par mail (Malika.Smail@loria.fr) et répondra à vos questions durant les dernières séances.

La **présence de tous.tes** aux **soutenances** est obligatoire.