Projet Décisionnel - Tabbone

Status	Not started

Sujet

Voici le sujet de mon projet :
Projet : La fouille de données en service du développement durable

Contexte : Big Datext, une 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 2 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 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.)

Important : 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.
La précision est calculée comme la moyenne des précisions des classifieurs dédiés à chaque classe (tronc, houppier, etc). Idem pour le rappel.

Défi 2 :
La seconde tâche, plus ouverte, vise à appliquer diverses techniques afin de mieux connaître et décrire l'état du "parc végétal" de Grenoble, de mieux comprendre son évolution et de 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, de réaliser des clusterings (avec un choix à faire des variables intéressantes à utiliser), rechercher des règles d'association parmi les variables.

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...

Déroulement du projet :
Le projet se fera par groupe, et chaque membre devra justifier d'au moins 20h de travail. Chaque groupe devra aborder les 2 parties du défi en utilisant un logiciel.

Un rapport devra être rédigé avec au moins la description de :

- l'exploration des données
- La préparation des données (toutes les variables sont elles intéressantes ? existe t il de fortes corrélations ? Quelle est la meilleure façon d'encoder numériquement chaque variable quand cela est nécessaire ?)
- des algorithmes appliqués aux données en justifiant le choix de ces algorithmes (au-moins 3 pour le défi 1)
- les évaluations rigoureuses et interprétation des résultats (est-ce que le modèle est meilleur que le classifieur baseline ? que disent les modèles en plus des métriques de performances ? sont ils en accord ? est-ce que ce sont les mêmes variables qui ont été détectées comme prédictives dans tous les modèles construits ?). Pour le défi 1, les critères d'évaluation à considérer incluent les performances, l'intelligibilité et la simplicité du modèle construit.

Soutenance : 24-25 mars

Rapport PDF : 21 mars au plus tard par mail à Tabbone