ENONCE DE PROJET ©

Matière : Fouille de données **Enseignants :** M. Taoufik Ben Abdallah

Discipline : 2ème année Génie Informatique Mme. Rahma Boujelben

Année Universitaire : 2023-2024 / S1

Afin de proposer des offres de contrats d'assurance habitation personnalisées, une assurance souhaite intégrer dans son système un modèle qui permet de prédire si un bâtiment aura un accident pendant la période d'assurance. Ainsi, nous lançons cette compétition afin de lui proposer le modèle de prédiction gagnant. Pour cela, nous mettons à votre disposition le dataset **`train_insurance.csv`**. Il est composé de **`5012`** observations décrites par **`12`** **attributs descripteurs** et une **variable classe** **`claim.`**

Voici une brève description des différentes variables :

- **`Customer Id`** : Le numéro d'identification du bénéficiaire
- **`YearOfObservation`** : L'année d'observation de l'état du bâtiment
- **`Insured_Period`** : La période d'assurance (`1` : `1` ans, `0.5` : `6` mois)
- **`Residential`** : Le bâtiment est-il résidentiel ? (**`1`** : oui, `0` : non)
- **`Building_Painted`** : Le bâtiment est-il peint ? (**`N`** : oui, **`V`** : non)
- **`Building_Fenced`** : Le bâtiment est-il clôturé ? (**`N`** : oui, **`V`** : non)
- **`Garden`** : Le bâtiment a-il un jardin ? (**`V`** : oui, **`O`** : non)
- **`Settlement`** : La zone du bâtiment. (**`R`** : zone rurale, **`U`** : zone urbain)
- **`Building Dimension`** : La taille du bâtiment en m^2
- **`Building_Type`** : Le type de bâtiment (**`'Fire-resistive', 'Non-combustible', 'Ordinary', 'Wood-framed'`**)
- **`NumberOfWindows`** : Le nombre de fenêtres du bâtiment (**without** dans le cas de **`0`** fenêtre)
- **`Geo Code`** : Le code géographique du bâtiment assuré
- **`Claim`** : La variable classe (**oui** si **le bâtiment a au moins une réclamation pendant la période d'assurance**, et **non** si **le bâtiment n'a pas eu de réclamation pendant la période d'assurance**)

Le **Tableau 1** montre un extrait du jeu de données **`train_insurance.csv`**

Customer Id	YearOfObservation	Insured_Period	Residential	Building_Painted	Building_Fenced	Garden	Settlement	Building Dimension	Building_Type	NumberOfWindows	Geo_Code	Claim
H13501	2012	1.0	1	N	V	V	U	1240.0	Wood- framed	without	75117	**non**
H14962	2012	1.0	0	N	V	V	U	900.0	Non-combustible	without	62916	**non**
H17755	2013	1.0	1	V	N	O	R	4984.0	Non-combustible	4	31149	**oui**
H13369	2016	0.5	0	N	V	V	U	600.0	Wood-framed	without	6012	**oui**
H12988	2012	1.0	0	N	V	V	U	900.0	Non-combustible	without	57631	**non**
H3052	2016	0.5	0	N	V	V	U	2675.0	Ordinary	without	38185	**non**

Tableau 1 : Extrait du jeu de données « `train_insurance.csv` »

NB. Le fichier « `test_insurance.csv` » représente le jeu de données de test. Il comporte `2147` observations, et sera utilisé pour l'évaluation et la validation des modèles générés.

Travail à faire :

1/ Analyser et visualiser les données

2/ Nettoyer, **si necessaire**, les données

3/ Sélectionner, **si necessaire**, les descripteurs les plus discriminants

4/ Encoder les données et générer un ou plusieurs modèle(s) de prédiction en appliquant des **techniques d'apprentissage supervisée** (`DecisionTreeClassifier, SVC, MLPClassifier GradientTreeBoosting, RandomForestClassifier,` etc.)

5/ Évaluer les performances du/des modèle(s) obtenu(s). **Interpréter les résultats**

Bon Travail ♣