DÉTECTION DE FRAUDES DANS LE DOMAINE DE L'ASSURANCE

Ce projet peut être réalisé en monôme ou en binôme. L'objectif de cette étude est de construire un modèle de prédiction des déclarations frauduleuses des clients d'une société d'assurance.

Ensembles de données

Les deux ensembles de données concernent les mesures qu'entreprend une société d'assurance pour détecter et prédire les cas de déclarations frauduleuses d'accidents :

- Le fichier `Data_Projet_1.csv` contient des informations financières et démographiques concernant 1100 déclarations d'accidents, avec pour chacune l'information s'il s'agissait d'une déclaration frauduleuse ou non.
- Le fichier `Data_Projet_1_New.csv` contient les informations sur 200 nouvelles déclarations pour lesquels la société souhaite prédire s'il y a un risque de fraude.

Caractéristiques des données :

- Instances : chaque instance correspond à une déclaration d'accident
- Nombre de variables : 12
- Séparateur de colonnes : , (virgule)
- Séparateur de décimales : . (point)
- Variable de classe : `FRAUDULENT`

Le dictionnaire des données ci-dessous décrit pour chacune des 12 variables son nom, son type, sa description et son domaine de valeurs.

Dictionnaire des données

Variable	Type	Description	Domaine de valeurs
`CLAIM_ID`	Numérique	Numéro unique d'identification de la déclaration	`[26832, 99775483]`
`CUSTOMER_ID`	Numérique	Numéro d'identification de l'assuré	`[154557, 99961993]`
`AGE`	Numérique	Age de l'assuré en nombre d'années	`[18, 79]`
`GENDER`	Catégoriel	Genre de l'assuré	`Male, Female`
`INCIDENT_CAUSE`	Catégoriel	Cause déclarée de l'accident	`Driver error, Natural causes, Other causes, Other driver error`
`DAYS_TO_INCIDENT`	Numérique	Délai écoulé en nombre de jours depuis l'accident	`[2, 14991]`
`CLAIM_AREA`	Catégoriel	Type de déclaration	`Auto, Home`
`POLICE_REPORT`	Catégoriel	Un rapport de police a-t-il été rédigé	`Yes, No, Unknown`
`CLAIM_TYPE`	Catégoriel	Type de déclaration d'accident	`Injury only, Material and injury, Material only`
`CLAIM_AMOUNT`	Numérique	Montant des dégâts déclarés	`[1000, 47748]`
`TOTAL_POLICY_CLAIMS`	Numérique	Nombre total de déclarations d'accidents de l'assuré	`[1, 8]`
`FRAUDULENT`	Catégoriel	La déclaration est-elle frauduleuse ?	`Yes, No`

Fichiers de données

Fichier	Nbr instances	Classe?	Remarques
`Data_Projet_1.csv`	1100	Oui	Instances dont la classe réelle est connue
`Data_Projet_1_New.csv`	200	Non	Instances à prédire (classe inconnue)

Objectifs du projet

L'objectif est la création d'un modèle de prédiction du risque de déclaration frauduleuses par les assurés et

son application aux nouvelles déclarations (instances à prédire). On souhaite donc utiliser les techniques de classification afin de générer un modèle de prédiction de la classe des déclarations :

- `FRAUDULENT = Yes` (positif)
- `FRAUDULENT = No` (négatif)

Plusieurs classifieurs seront générés et testés en appliquant les différentes méthodes de classification et en ajustant les paramètres afin d'optimiser les résultats. Seul le classifieur le plus performant sera conservé sachant que l'**on souhaite avant tout minimiser les risques financiers en évitant de prédire comme non-frauduleuse une déclaration effectivement frauduleuse**.

Le classifieur sélectionné sera ensuite appliqué à l'ensemble de données à prédire afin de prédire pour chaque déclaration si elle est susceptible d'être frauduleuse (classe `FRAUDULENT = Yes`) ou non (classe `FRAUDULENT = No`).

Afin d'évaluer les classifieurs générés, vous définirez un ou des critère(s) (basés sur les taux de succès/échecs, la matrice de confusion ou les mesures d'évaluation par exemple) **en fonction des objectifs de l'application décrits ci-dessus**. Vous comparerez les résultats du test des classifieurs générés sur l'ensemble de test selon ce(s) critère(s) afin d'identifier le plus pertinent.

Processus d'analyse

Le processus général pour cette analyse suivra les étapes suivantes :

- Exploration et visualisation des données.
- Pré-traitement des données.
- Extraction de clusters à partir des données.
- Définition de la méthode d'évaluation des classifieurs.
- Définition des données d'apprentissage et de test.
- Construction et évaluation des classifieurs.
- Choix du classifieur le plus performant.
- Application du classifieur aux données à prédire.

Référez-vous aux méthodes appliquées durant les séances de Travaux Dirigés pour chacune de ces étapes.

Rapport de projet

Vous devez rendre comme rapport de votre projet :

- Un rapport au **format .pdf** décrivant tous les traitements que vous avez effectué et les résultats obtenus :
  - Exploration des données et interprétation des résultats (relations notables, problèmes, variables les plus utiles pour la prédiction de la classe, etc.).
  - Pré-traitements appliqués aux données si besoin (sélection des variables, transformation des valeurs, etc.).
  - Description des configurations algorithmiques utilisées pour le clustering des données (algorithmes et paramétrages) et évaluation des clusters obtenus vis-à-vis de la classe `FRAUDULENT = Yes` ou `FRAUDULENT = No` (i.e., proportion de chaque classe dans le cluster) afin de générer des clusters correspondants chacun aussi majoritairement que possible à l'une des deux classes.
  - Description des meilleurs clusters obtenus avec pour chacun la liste des caractéristiques spécifiques à ce cluster : l'objectif est d'identifier pour chacune des deux classes des sous-groupes (clusters) de clients correspondant de la même classe et ayant des caractéristiques communes (par exemple même catégorie d'âge, même nombre d'enfants, etc.) qui sont spécifiques à ce cluster, c-à-d qui sont différentes de celles autres clusters (sous-groupes de clients) correspondant à la même classe.
  - Définition de la méthode d'évaluation des classifieurs (taux de succès/échecs, matrices de confusion, mesures d'évaluation, etc.) pour la sélection du classifieur le plus pertinent en fonction des objectifs.
  - Description de la méthode de création des données d'apprentissage et de test : techniques utilisées (partitionnement, échantillonnage, etc.) et leur paramétrage(s), etc..
  - Description des configurations des classifieurs générés (algorithmes et paramétrages) et évaluation de leurs performances selon la méthode d'évaluation définie précédemment. Vous indiquerez quel(s) est(sont) le(s) classifieur(s) donnant les meilleurs résultats selon cette méthode d'évaluation.
  - Description du classifieur sélectionné (type de modèle, algorithme, paramétrage, matrice des coûts, etc.) et éventuellement de sa structure (dimensions de l'arbre de décision, nombre de règles de classification, etc.) ; C'est à dire tous les éléments qui vous paraissent utiles pour décrire sa structure, sa complexité et sa pertinence.

- Description résumée des résultats de l'application du classifieur sélectionné à l'ensemble de données à prédire (répartition des classes, probabilités minimales, maximales et moyennes associées à chacune des classes, etc.).
- Conclusion résumant vos autres observations sur cette application et les résultats, les difficultés rencontrées, etc..

➢ Le fichier au **format R** contenant vos commandes R.

➢ Un fichier au **format .csv** contenant les résultats de l'application du classifieur sélectionné à l'ensemble à prédire afin de fournir une prédiction de la classe pour chacun des nouveaux clients.

Le résultat doit être représenté sous forme d'un tableau avec sur chaque ligne :

- Le numéro d'identification du client.
- La classe prédite pour ce client.
- La probabilité associée à la prédiction de cette classe.

Indiquez votre(vos) nom(s) et prénom(s) sur la première page du rapport.