# Prédiction de Défauts de Payement

Ce projet peut être réalisé en monôme ou en binôme. Vous utiliserez pour le réaliser les ensembles de données `projet.csv` et `projet_new.csv` disponibles sur le LMS UniCA.

## Ensembles de données

Ces deux ensembles de données concernent les mesures qu'entreprend une banque pour réduire le taux de défauts de payement des remboursements d'emprunts. Le fichier `projet.csv` contient des informations financières et démographiques concernant 6000 clients ayant déjà effectué un emprunt, avec pour chacun l'information sur un défaut de payement survenu ou non. Le fichier `projet_new.csv` contient les informations sur 500 clients pour lesquels la banque souhaite prédire s'il y a un risque de défaut de payement pour l'octroi d'un emprunt.

Caractéristiques des données :

- Instances : chaque instance correspond à un client identifié par son numéro
- Nombre de variables : 11
- Séparateur de colonnes : , (virgule)
- Séparateur de décimales : . (point)
- Variable de classe : `defaut`
- Variables avec valeurs manquantes : `age`, `adresse`

Le dictionnaire des données ci-dessous décrit pour chacune des 11 variables son nom, son type (entier, réel, booléen, catégoriel ou ordinal), sa description, son domaine de valeurs (liste de valeurs ou nombres minimal et maximal) et le codage des valeurs manquantes s'il y en a.

***Dictionnaire des données***

| Variable | Type | Description | Domaine de valeurs | Valeurs manquantes |
|---|---|---|---|---|
| `client` | Entier | Numéro d'identification du client | `[1201, 8500]` | |
| `age` | Entier | Age en nombre d'années | `[18, 999]` | `999` |
| `education` | Ordinal | Niveau d'éducation relativement au baccalauréat | `Niveau bac`<br>`Bac+2`<br>`Bac+3`<br>`Bac+4`<br>`Bac+5 et plus` | |
| `emploi` | Entier | Nombre d'années avec l'employeur actuel | `[0, 63]` | |
| `categorie` | Entier | Catégorie bancaire | `[12, 12]` | |
| `adresse` | Entier | Nombre d'années à l'adresse actuelle | `[0, 999]` | `999` |
| `revenus` | Réel | Revenus du foyer en milliers de $ | `[12.3, 2461.7]` | |
| `debcred` | Réel | Ratio Débit/Crédit (x100) | `[0.08, 44.62]` | |
| `debcarte` | Réel | Débit carte de crédit en milliers de $ | `[0.005, 139.580]` | |
| `autres` | Réel | Autres dettes en milliers de $ | `[0.009, 416.517]` | |
| `defaut` | Booléen | Un défaut de payement a-t-il eu lieu ? | `Oui`<br>`Non` | |

***Fichiers de données***

| Fichier | Nombre d'instances | Classe? | Remarques |
|---|---|---|---|
| `projet.csv` | 6000 | Oui | Instances dont la classe réelle est connue |
| `projet_new.csv` | 500 | Non | Instances à prédire |

## Objectifs du projet

L'objectif est la création d'un modèle de prédiction du risque de défaut de payement pour les clients et son application aux instances à prédire. On souhaite donc utiliser les techniques de classification afin de générer un modèle de prédiction de la classe des clients :

- `defaut = Oui` (risque positif)
- `defaut = Non` (risque négatif)

Plusieurs classifieurs seront générés et testés en appliquant les différentes méthodes de classification et **en ajustant leurs 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 d'accorder un emprunt à tort**.

Le classifieur sélectionné sera ensuite appliqué à l'ensemble de données à prédire afin de prédire pour chaque client s'il est susceptible d'avoir un défaut de payement (classe `defaut = Oui`) ou non (classe `defaut = Non`).

Afin d'évaluer les classifieurs générés, vous définirez un ou des critère(s) -- basé(s) sur les taux de succès/échecs, la matrice de confusion, les mesures d'évaluation ou/et les courbes ROC, par exemple -- en fonction des objectifs de l'application décrits ci-dessus. Vous comparerez les résultats obtenus sur l'ensemble de test pour chacun des classifieurs générés selon ces critères afin d'identifier le plus pertinent.

L'analyse exploratoire des données devra permettre d'identifier d'éventuels problèmes dans les données (déséquilibre très important des classes, erreurs dans les données, etc.), d'identifier d'éventuels pré-traitement des données à effectuer (corrections des problèmes, typage de variables, etc.) et d'appréhender l'utilité de chacune des variables prédictives pour prédire les classes.

Le clustering des données devra permettre d'identifier si possible des groupes de clients similaires (selon leurs caractéristiques d'âge, niveau d'éducation, etc.), chaque groupe correspondant très majoritairement à l'une des deux classes. Si de tels groupes sont obtenus, une analyse comparative des groupes (moyenne des âges des clients du groupe, niveau d'éducation le plus fréquent dans le groupe, etc.) devra être faite afin de mieux comprendre combiens de groupes il existe pour chaque classe `defaut = Oui` et `defaut = Non`, et comprendre ce qui distingue ces classes (différentes valeurs moyennes d'âge, de niveaux d'éducation, etc.).

## Processus d'analyse

Le processus général pour cette analyse suivra les étapes suivantes :

- Analyse exploratoire des données.
- Pré-traitement des données.
- Clustering des données.
- Définition de la méthode d'évaluation des classifieurs.
- Définition des données d'apprentissage et de test (à partir du fichier `projet.csv`).
- Construction et évaluation des classifieurs.
- Choix du classifieur le plus performant.
- Application du classifieur le plus performant aux données à prédire (instances du fichier `projet_new.csv`).

Référez-vous aux méthodes appliquées durant les tutoriels pour chacune de ces étapes.

## Rapport de projet

Vous devez déposer sur le LMS UniCA dans la boîte de dépôt *Rapport de 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 résultats pour la distinction des classes.
  - Caractérisation des clusters du meilleur clustering obtenu, avec pour chaque cluster une description des caractéristiques représentatives du cluster, c-à-d qui permettent de différencier les clusters : descriptions des valeurs des variables numériques et répartitions des valeurs des variables catégorielles pour les instances du cluster.
  - 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 leur

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, etc.) et évaluations détaillée des résultat obtenus lors du test de ce classifieur pour un maximum de critères (matrice de confusion, taux de succès, mesures de Précision, Rappel, Spécificité, Taux de Vrais Négatifs, etc.).
  - Résumé 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..

- 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.
- Le **fichier de commandes R** correspondant à toutes les opérations que vous avez réalisé pour votre projet. Commenter un minimum votre fichier (une courte phrase pour décrire ce qu'un bloc d'instructions fait par exemple, inutile de commenter chaque ligne).

## Consignes

- Mentionnez dans votre rapport tout ce que vous avez testé (méthodes, algorithmes, paramétrages, visualisations, comparaisons, etc.), même si cela n'a pas donné de résultat probant ou utile. Le fait qu'un test réalisé ne donne pas de résultat pertinent est une connaissance utile pour la compréhension de l'application, des données, etc. Inutile d'entrer dans le détail sur ces points toutefois, soyez concis et bref sur ceux-ci s'il y en a.
- Indiquez votre(vos) nom(s) et prénom(s) sur la première page du rapport.
- Nommez vos fichier avec les NOMS et Prénoms des membres de votre groupe.
  Par exemple :
  - PASQUIER_Nicolas_DUPOND_Jean.pdf
  - PASQUIER_Nicolas_DUPOND_Jean.csv
  - PASQUIER_Nicolas_DUPOND_Jean.R
- Déposez les 3 fichiers dans la boîte de dépôt sur le LMS UniCA **au plus tard le dimanche 1er décembre 2024**.