# Projet Applicatif - Prédiction de survie d'une startup en France

Membres du groupe : Elizabeth Wolfsohn, Etienne De Poix et Amele Olivia ADAMAH

Résumé du projet: Nous nous intéressons à un problème de classification dans lequel nous proposons de construire un modèle de prédiction permettant de déterminer si une startup en France "survit" après quelques années d'existence. Nous définirons la survie comme le fait qu'une entreprise soit active (statut administratif). Le projet couvrira une petite période d'années (ex: 2019-2021), selon le nombre de données accessibles, et portera principalement sur des startups dans des secteurs innovants (ex: santé, finance, tech, défense, etc…).

Objectif et problématique : L'objectif est de prédire à partir des caractéristiques financières et administratives d'une jeune entreprise si elle sera encore active après quelques années. Peut-on prédire sa pérennité sur la base de données financières et structurelles ? Quels facteurs influencent le plus la survie d'une startup ?

Données Utilisées

1. **Base Sirene des entreprises et de leurs établissements, INSEE**

Base Sirene des entreprises et de leurs établissements (SIREN, SIRET)

**Données intéressantes:**

StockUniteLegale: siren, date de création de l'entreprise, état actif ou non, numéro de secteur, tranche d'effectif

StockEtablissement: données liées aux localisations des établissements

2. **Données financières - Signaux Faibles**

Données financières détaillées des entreprises (format parquet)

contient les bilans financiers des entreprises par an (détails des données financières précisés avec des codes à deux lettres identifiables grâce au document 2050-liasse_5013.pdf)

**Données intéressantes**: siren, chiffre d'affaires, dettes, résultat de l'exercice (bénéfices ou pertes)

Approche méthodologique : préparation du jeu de données, filtrage des entreprises sur une petite période d'années, sélection des entreprise dans les secteurs innovants, fusion des datasets via la clé SIREN, création de nouvelles features : label final (1 = survie, 0 = arrêt), ancienneté, etc…

Méthodes de machine learning

- **Régression logistique:** bien pour une première approche (simple et interprétable)
- **Réseaux de neurones** : qui est flexible et capable de détecter les relations complexes entre les features, et pour avoir un meilleur apprentissage, surtout avec notre jeu de données assez volumineux.
- **Random Forest** : pour sa robustesse, sa précision, et pour identifier les variables les plus importantes.
- **Arbre de décision:** facilité pour la visualisation