Avril 2025

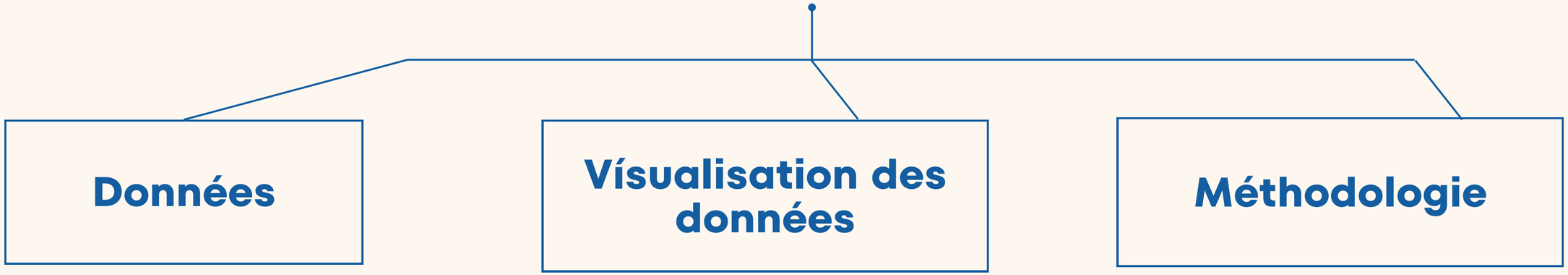
Rennes

Présentation

Introduction à la science des données

Introduction du rapport

Nous considérons un ensemble de données provenant d'une base de crédit appelée credit_risk_dataset. Cet ensemble contient des informations sur des emprunteurs, telles que l'âge, le revenu, le type de logement, la durée d'emploi, ainsi que des détails sur les prêts comme le montant, le taux d'intérêt ou encore l'objectif du prêt. Ce dataset est utilisé pour analyser le risque de défaut de paiement et pour entraîner des modèles de machine learning capables de prédire la probabilité de non-remboursement.



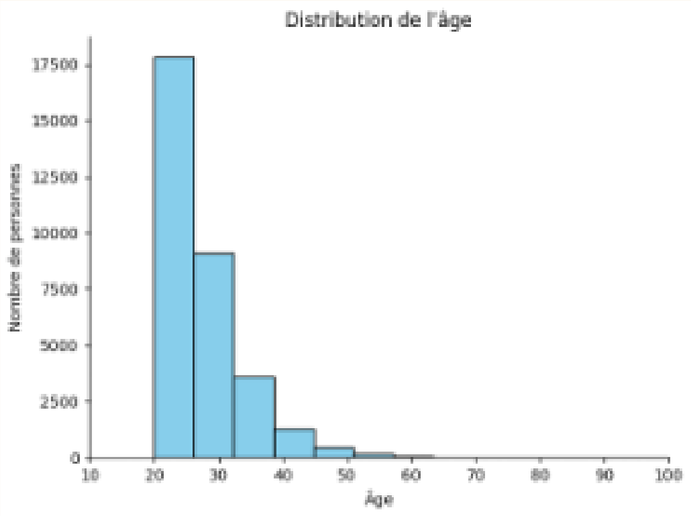
Données

Nom de la variables	Description	Type
person_age	Âge	int64
person_income	Revenue annuel	int64
person_home-ownership	Accès à la propriété	object
person_emp_length	Durée de l'emploi(en années)	float64
loan_intent	Intention de prêt	object
loan_grade	Catégorie de prêt	object
loan_amnt	Montant du prêt	int64
loan_int_rate	Taux d'intérêt	float64
loan_status	Statut du prêt(0 n'est pas par défaut, 1 est par défaut	int64
loan_percent_income	Pourcentage de revenu	float64
cb_person_default_on_file	Défaut historique	object
cb_preson_cred_hist_length	Durée de l'historique de crédit	int64

nombre de variables: 12

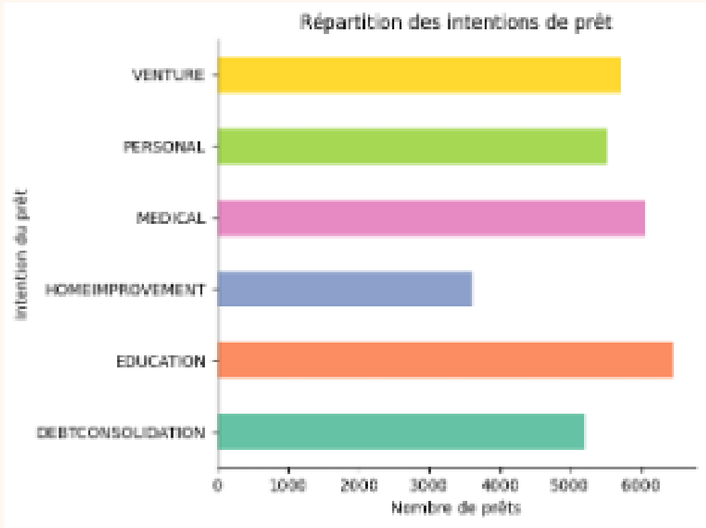
nombre d'observations: 32581

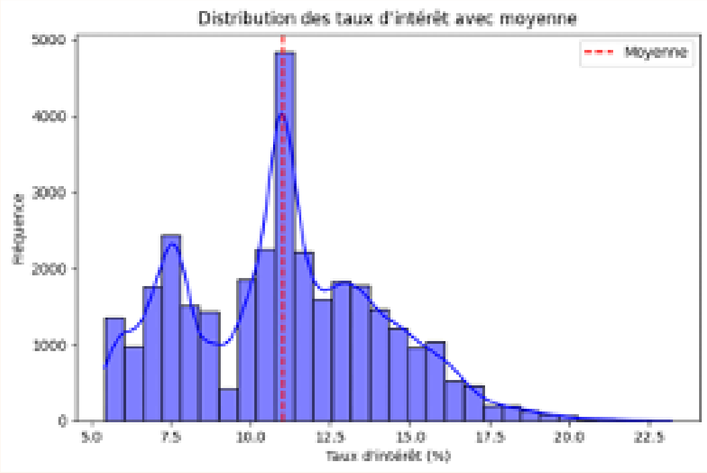
Visualisation des données



<- Pic chez les 20-49 ans, risque potentiel lié aux jeunes emprunteurs (instabilité financière)

Priorité à l'éducation et au médical ->





<- Taux bas dominants (10-12.5%)

Méthodologie

Notre approche combine méthodes exploratoires et prédictives pour analyser le risque de crédit :

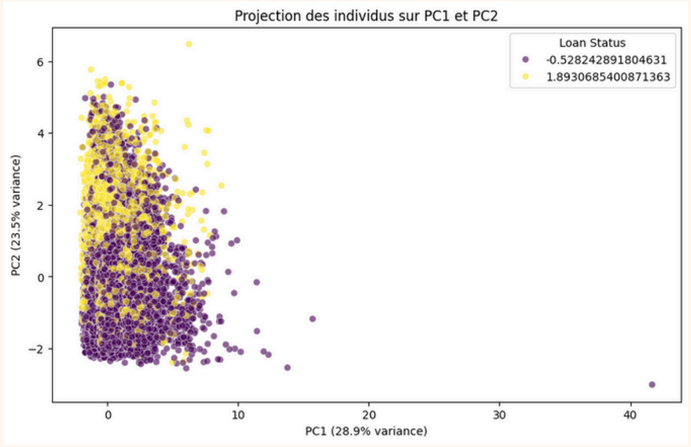
ACP	**Réduction de dimension et identification des variables influentes.**
K-means	**Segmentation des emprunteurs selon leur profil de risque**
Permutation Importance	**Hiérarchisation des facteurs clés**
Comparaison de modèles (KNN, Forêt Aléatoire, Régression Linéaire)	**Évaluation de la performance prédictive pour cibler le meilleur algorithme.**

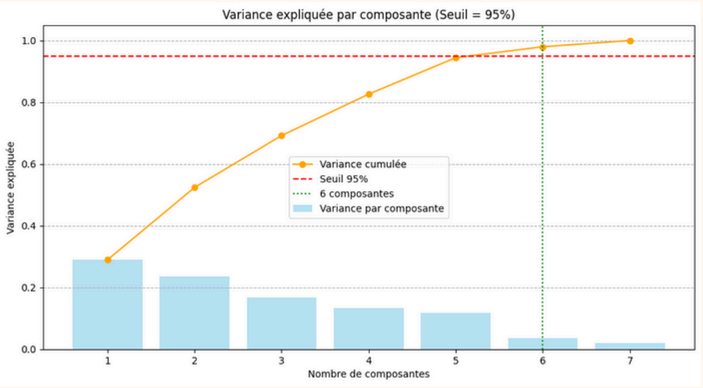
Objectif : Identifier les déterminants du défaut et optimiser la prédiction du risque.

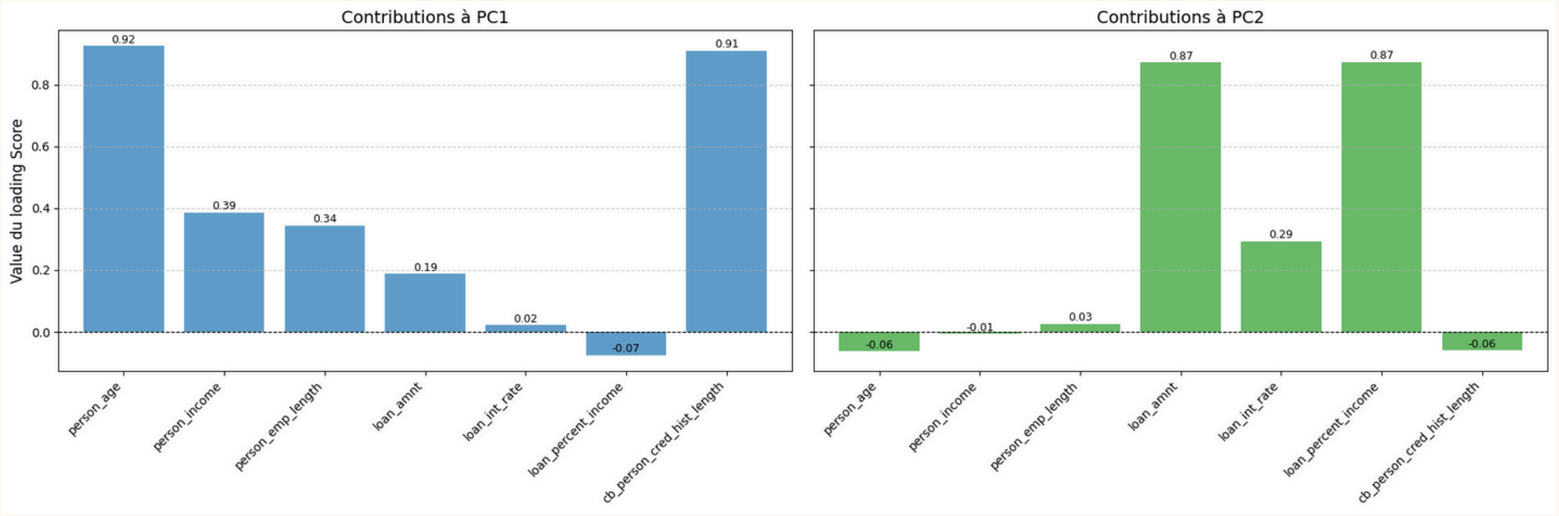
ACP

Les axes PC1 et PC2 offrent une distinction claire entre les différents profils d'emprunteurs, facilitant l'identification des profils à risque.

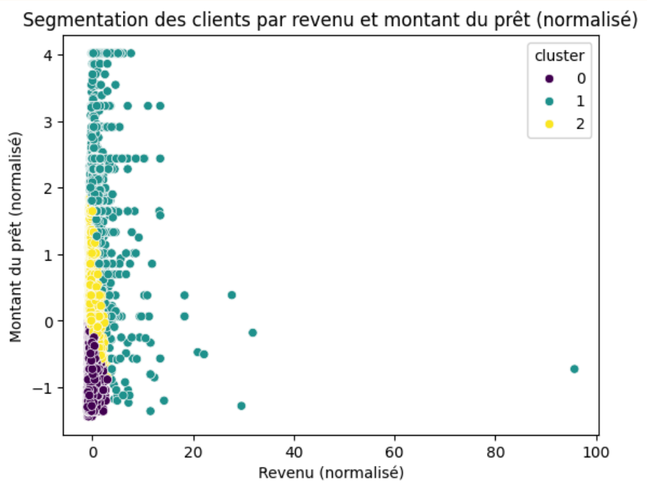
La projection des individus confirme la capacité de l'ACP à former des clusters selon le statut du prêt (remboursé ou en défaut).







K-Means



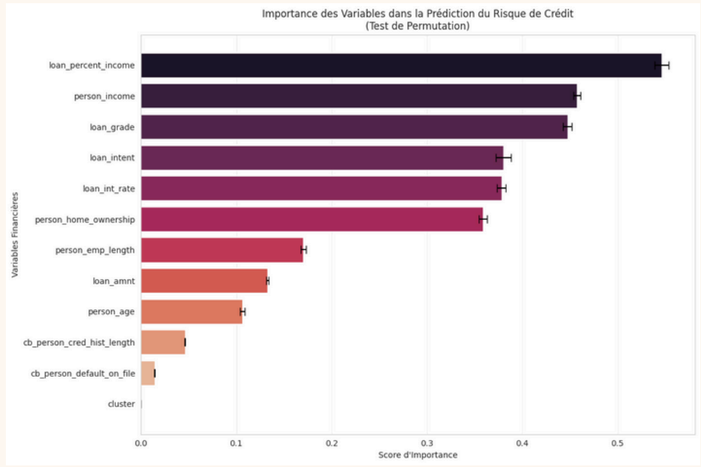
Segmentation clients (K-means)

- **3 clusters :**
  - **1 : Bas revenu / Petit prêt → Risque modéré (ratio élevé).**
  - **2 : Revenu moyen / Prêt moyen → Risque variable.**
  - **3 : Haut revenu / Gros prêt → Risque faible (solvabilité).**
- **Clusters distincts → Méthode robuste.**

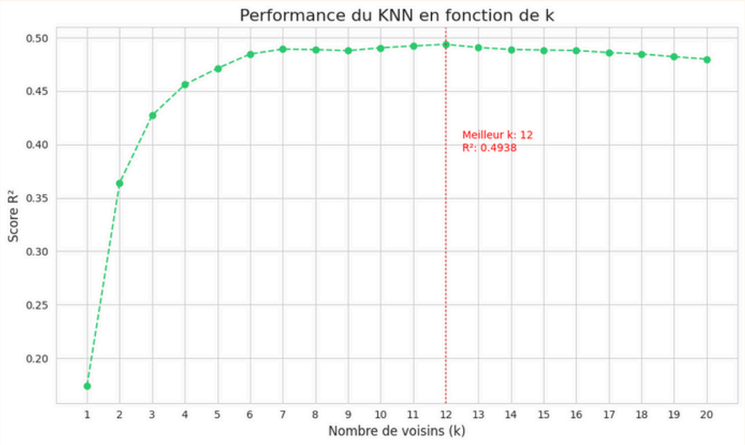
permutation_importance

Importance des variables (Permutation)

- **Top 3 :**
  - **loan_percent_income (Score ~0.5) → Ratio prêt/revenu critique.**
  - **person_income → Revenu influence la solvabilité.**
  - **loan_grade → Qualité du prêt prédictive.**
- **Historique crédit :**
  - **ch_person_default_on_file → Défauts passés augmentent le risque.**
  - **ch_person_cred_hist_length → Expérience crédit = risque réduit.**
- **Autres : loan_intent (objectif) et loan_int_rate (taux) moins impactants.**

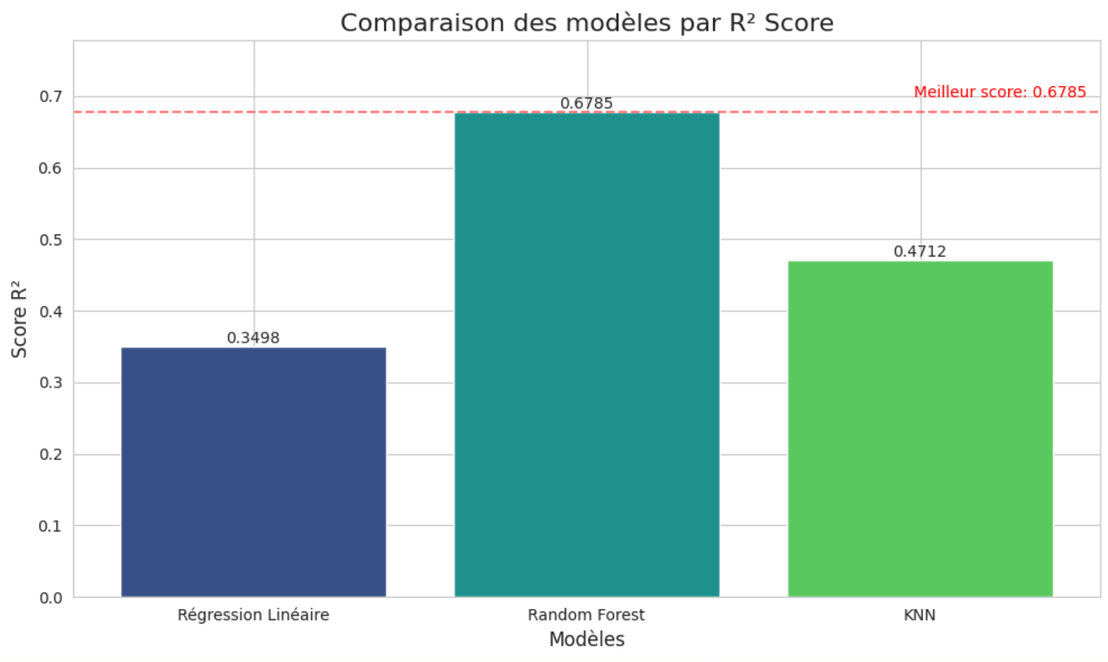


KNN



- **Meilleur k : 12 → Score R^2 = 0.49 (performance maximale).**
- **Tendance :**
  - **k faible (1-5) : Surapprentissage (score instable).**
  - **k élevé (>15) : Sous-apprentissage (score dégradé).**

Comparaison des 3 modèles par R^2 score



Comparaison des modèles (R^2)

- **Forêt Aléatoire : 0.68 → Meilleure performance (relations complexes bien capturées).**
- **KNN : 0.47 → Limité par le bruit ou le choix de k.**
- **Régression Linéaire : 0.35 → Inadaptée aux non-linéarités**

Conclusion :

→ La Forêt Aléatoire est retenue pour la prédiction du risque.

Conclusion

1. Capacité de prédiction du statut du prêt
   - Forêt Aléatoire : Modèle le plus performant ($R^2 = 0.68$), adapté aux relations complexes (ex. interactions entre variables).
   - KNN : Score intermédiaire ($R^2 = 0.47$), améliorable par optimisation des hyperparamètres.
   - Régression Linéaire : Peu efficace ($R^2 = 0.35$) en raison de la non-linéarité des données.
2. Importance des variables
   - Ratio prêt/revenu : Variable la plus influente (permutation importance).
   - Revenu de l'emprunteur : Solvabilité directement liée au risque.
   - Historique de crédit : Défauts passés augmentent significativement le risque.
   - Grade du prêt : Grades inférieurs (C-G) associés à des taux variables et risques élevés.
3. Choix des modèles
   - Forêt Aléatoire retenue pour :
     - Sa capacité à gérer les non-linéarités et interactions.
     - Sa robustesse face au bruit (ex. données mal étiquetées).
   - KNN et Régression Linéaire écartés pour des performances inférieures.

Merci pour votre attention