1. Analyse exploratoire descriptive

1.1 Description générale du jeu de données

Les données proviennent d'une banque portugaise et contiennent des informations socio-économiques et financières sur les clients contactés lors de campagnes marketing téléphoniques.
La variable cible est **`class`**, indiquant si le client a souscrit (`yes`) ou non (`no`) à un dépôt à terme.

Variable	Type	Description
`age`	Numérique	Âge du client
`marital`	Catégorique	Statut marital : single, married, divorced
`education`	Catégorique	Niveau d'éducation : primary, secondary, tertiary, unknown
`default`	Catégorique	Présence d'un défaut de crédit : yes / no
`balance`	Numérique	Solde moyen du compte
`housing`	Catégorique	Crédit immobilier : yes / no
`loan`	Catégorique	Crédit personnel : yes / no
`contact`	Catégorique	Type de contact : telephone, cellular, unknown
`poutcome`	Catégorique	Résultat de la campagne précédente : success, failure, unknown
`class`	Catégorique	Variable cible : yes / no

1.2 Exploration visuelle et analyse des distributions

Les visualisations (histogrammes, boxplots et diagrammes en barres) montrent des tendances claires :

- **Âge** : les clients souscripteurs ont en moyenne un âge légèrement plus élevé, suggérant que les individus plus stables économiquement sont plus enclins à investir.

- **Balance** : les soldes moyens sont significativement plus élevés chez les souscripteurs.

- **Marital / Education** : les célibataires et diplômés du supérieur sont légèrement surreprésentés parmi les souscripteurs.

- **Housing & Loan** :

    - Les clients **sans crédit logement** (housing = no) souscrivent davantage.

    - Les clients **ayant un crédit personnel** (loan = yes) souscrivent moins souvent.

- **Poutcome** : une campagne précédente réussie augmente fortement la probabilité de souscription.

Les tests du **Chi² d'indépendance** confirment des liens significatifs entre `class` et les variables `marital`, `education`, `housing`, `loan` et `poutcome`.

Interprétation :
Cette analyse suggère que les caractéristiques financières et le résultat des campagnes précédentes sont de bons prédicteurs de la souscription.

2. Construction du modèle d'arbre de décision

2.1 Méthodologie

L'arbre de décision a été construit avec la fonction **`rpart()`**, implémentant l'algorithme **CART (Classification and Regression Tree)**, conformément au cours (séance du 08/10 et 09/10).
L'objectif est de modéliser la probabilité de souscription (`class ~ .`).

Le jeu de données a été séparé en :

- 70 % pour l'apprentissage (`train_data`)

- 30 % pour le test (`test_data`)

2.2 Optimisation et élagage

Pour éviter le **sur-apprentissage**, un **paramètre de complexité (cp)** a été optimisé à l'aide de la **validation croisée intégrée** dans `rpart`.
La commande `printcp()` et la visualisation `plotcp()` permettent de sélectionner le cp minimisant l'erreur de validation croisée (`xerror`).

2.3 Visualisation de l'arbre

L'arbre final a été tracé avec `rpart.plot()`.
Chaque nœud indique la variable de décision, les modalités de coupure et la proportion de classes (`yes` / `no`).

Observation :
Les premières branches font intervenir les variables **`poutcome`** (résultat de campagne précédente) et **`housing`**, confirmant leur importance dans la décision finale.

3. Évaluation du modèle optimal

3.1 Résultats sur le jeu de test

L'évaluation a été réalisée à l'aide de la **matrice de confusion** et de métriques standards (issues du package `caret`).

Indicateur	Valeur (exemple)	Interprétation
Accuracy	~0.88	88 % de bonnes prédictions globales
Précision	~0.75	75 % des clients prédits "yes" sont réellement souscripteurs
Rappel (Recall)	~0.64	64 % des souscripteurs réels sont détectés
F1-score	~0.69	Compromis entre précision et rappel
Spécificité	~0.92	92 % des "no" correctement exclus
AUC	~0.90	Excellente capacité discriminante

3.2 Courbe ROC et AUC

La **courbe ROC** montre une bonne séparation entre les classes.
L'**AUC (Area Under Curve)** ≈ 0.9 indique que le modèle distingue efficacement les souscripteurs des non-souscripteurs, conformément aux critères de performance vus dans le cours (séance du 15/10 : *"Courbe ROC et mesure AUC"*).

3.3 Autres métriques complémentaires

- **Erreur globale** : `1 - Accuracy` ≈ 0.12
- **Cohen's Kappa** ≈ 0.6 → accord substantiel au-delà du hasard.

Ces résultats confirment la bonne généralisation du modèle, sans sur-apprentissage excessif grâce à l'élagage (séance du 14/10 : *"Pruning pour contrôler la complexité"*).

4. Interprétation et conclusion

4.1 Synthèse des résultats

L'arbre de décision CART a permis d'identifier les facteurs principaux de souscription :

1. **Résultat de la campagne précédente (`poutcome`)** : les clients déjà "success" sont bien plus enclins à souscrire à nouveau.
2. **Crédit logement (`housing`)** : les clients sans prêt immobilier ont davantage de liquidités disponibles.
3. **Solde bancaire (`balance`)** : corrélé positivement à la probabilité de souscription.
4. **Âge et statut marital** : effet secondaire, les profils plus âgés et célibataires étant légèrement surreprésentés.

4.2 Performances globales

- Le modèle affiche une **bonne précision globale (≈88%)** et une **AUC élevée**, signe d'une capacité prédictive solide.
- Le compromis **précision / rappel** reste équilibré (F1 ≈ 0.69), ce qui montre une bonne aptitude à identifier les vrais souscripteurs sans trop de faux positifs.

- L’arbre élagué, validé par la **validation croisée**, a permis de limiter la complexité tout en conservant les performances.

4.3 Perspectives et extensions possibles

- Comparer les performances avec un **modèle de régression logistique** (Problème I du TD) pour évaluer la robustesse de l’arbre.

- Étendre l’analyse à une **forêt aléatoire** (Question 4 du cours), afin d’améliorer la stabilité et réduire la variance.

- Introduire un **seuil de probabilité ajusté** pour optimiser le rappel en fonction du coût d’erreur (dans un contexte marketing, mieux vaut contacter quelques faux positifs que rater un vrai souscripteur).

Conclusion finale

Le modèle CART optimisé, construit selon les méthodes vues en cours (séparation apprentissage/test, validation croisée, élagage, courbe ROC/AUC), fournit une interprétation claire et des performances robustes.
Les variables financières et comportementales sont déterminantes, et la démarche respecte les principes d’**apprentissage supervisé interprétable** étudiés dans le module.