Comparaison des modèles :

Dans un des projets, nous devons prédire si un passager a **survécu** au naufrage du Titanic. L’objectif est de minimiser les **faux négatifs** (ne pas rater des survivants).

* **Précision (Precision)** mesure la proportion de prédictions positives correctes, mais elle ne prend pas en compte les faux négatifs.
* **L’Accuracy** peut être trompeuse si la classe majoritaire domine (ex. beaucoup plus de non-survivants que de survivants).
* **L’AUC-ROC** donne une vue d’ensemble, mais dans un contexte critique comme la survie, on préfère maximiser le rappel.
* **Le rappel (Recall)** est crucial ici, car il indique la capacité du modèle à détecter tous les survivants. Un rappel élevé signifie que peu de survivants sont classés à tort comme "décédés".

Dans le cadre de ce projet, notre objectif est de prévoir si un client va accepter une proposition bancaire. Le but est de réduire les faux négatifs (ne pas passer à côté de clients susceptibles d'accepter l'offre), car cela contribuerait à améliorer les campagnes publicitaires et à accroître le taux de conversion.  
  
 Importance des indicateurs de performance : La précision (Precision) évalue le rapport entre les prédictions positives justes et l'ensemble des prédictions positives, sans tenir compte des faux négatifs. Il peut y avoir des problèmes si nous favorisons la précision au lieu du rappel.  
Si la classe majoritaire prédomine, l'Accuracy peut induire en erreur (par exemple, s'il y a beaucoup plus de clients qui rejettent l'offre que de ceux qui l'acceptent). Si 90 % des clients disent non, un modèle qui prédit toujours « non » affichera une précision de 90 %, mais il sera sans utilité.

**Le rappel (Recall)** est crucial ici, car il indique la capacité du modèle à détecter tous les clients susceptibles d’accepter l’offre. Un rappel élevé signifie que **peu de clients intéressés sont classés à tort comme "non intéressés"**, ce qui maximise les opportunités commerciales et enifin L'AUC-ROC offre un aperçu du compromis entre la sensibilité et la précision, mais lorsque l'objectif est d'optimiser les conversions, le rappel prend une plus grande importance.

Classement des 7 algorithmes selon les métriques.

Lalgorithme avec le meilleur rappel ici est le Naive Bayes, suivi de la Regression Logistique, du Foret Aleatoire, de lArbre de decision, K-Nearest Neighbors, Support Vector Machine et Réseaux de neurones.

○ Comparaison entre les datasets (Titanic vs Bank Marketing).

**Analyse des tendances de performance**: Pour comparer les modèles sur les deux datasets, une première étape essentielle consiste à observer les tendances générales des performances. En analysant les moyennes des métriques clés (précision, rappel, AUC-ROC, etc.), on peut déterminer si les modèles performants sur le dataset Titanic conservent leur efficacité sur le dataset Bank Marketing. Par exemple, un modèle de régression logistique, bien adapté aux données tabulaires simples avec peu de variables comme Titanic, pourrait se révéler moins performant pour Bank Marketing, qui présente davantage de variables et une structure plus complexe. En revanche, des modèles tels que la Forêt Aléatoire ou les réseaux de neurones pourraient offrir de meilleures performances sur des données plus volumineuses et riches comme celles de Bank Marketing.

**Écart d'Accuracy entre les datasets** Lorsque l'Accuracy d'un modèle diffère fortement entre Titanic et Bank Marketing, cela révèle une dépendance significative à la nature des données. Une haute Accuracy sur Titanic et une faible sur Bank Marketing peut indiquer que le modèle fonctionne bien avec un faible nombre de variables mais éprouve des difficultés face à des datasets plus complexes et riches en caractéristiques. Cela reflète les différences dans la manière dont les modèles interprètent et exploitent les relations entre les variables. Par exemple, un arbre de décision excelle sur Titanic en identifiant des règles simples (comme l'influence de la classe et du sexe sur la survie), mais peut avoir des performances limitées sur Bank Marketing, où des relations plus complexes entre les variables nécessitent des approches plus élaborées.

Certains modèles peuvent **bien fonctionner sur un dataset mais échouer sur un autre**. Par exemple, **Naive Bayes** fonctionne souvent bien sur des datasets où les hypothèses d’indépendance entre les variables sont valides, ce qui peut être approximativement le cas pour Titanic. Cependant, sur Bank Marketing, où les variables sont plus interconnectées (ex: influence du taux Euribor, de la durée de l’appel, etc.), Naive Bayes peut être beaucoup moins performant.

L'analyse des performances des modèles sur les deux ensembles de données permet de distinguer les modèles qui présentent une bonne capacité de généralisation de ceux qui réagissent au type de données. Un modèle performant sur Titanic et Bank Marketing est sans doute solide et adaptable, alors qu'un modèle qui ne fonctionne bien que sur l'un des deux pourrait être trop particulier et nécessiterait des modifications.

Comparaison des méthodes de validation :

○ Observations sur les scores en split 70/30 vs cross-validation 5/7/10 folds.

* Les scores obtenus avec **CV (5 folds), CV (7 folds) et CV (10 folds)** sont très proches pour tous les modèles, ce qui montre que la **performance est stable** et ne dépend pas fortement de la manière dont les données sont divisées.
* Exemple : Régression Logistique : 0.797703 **(5 folds),** 0.796148 **(7 folds),** 0.792908 **(10 folds)**
* SVM : 0.831419 **(5 folds),** 0.834671 **(7 folds),** 0.834665 **(10 folds)**
* Cela indique que les modèles **généralisent bien** et ne sont pas excessivement sensibles aux variations des jeux d’entraînement et de test.
* En général, l’**accuracy du split 70/30 est légèrement supérieure** à celle obtenue en validation croisée.
* Exemple :
* **Régression Logistique** : CV (10 folds) = 0.792908, Split 70/30 = 0.809701
* **Forêt Aléatoire** : CV (10 folds) = 0.804378, Split 70/30 = 0.783582
* Cette différence peut s’expliquer par le fait que le split 70/30 teste le modèle sur **un seul ensemble de test**, tandis que la validation croisée répartit les données de manière plus équilibrée, limitant ainsi la dépendance à un jeu de test particulier.

○ Discussion sur la variance des scores et la robustesse de l’estimation.

La validation croisée démontre une **faible variance**, témoignant de la **fiabilité et de la cohérence** des estimations de performance des modèles. Parmi eux, la **Régression Logistique ,le SVM** et le **K-Nearest Neighbors** se distinguent par leur **robustesse**, affichant des scores remarquablement **stables** quelle que soit la méthode de validation employée. À l’inverse, l’**arbre de décision** révèle une **sensibilité accrue** aux variations de l’ensemble de test, avec des fluctuations plus marquées entre le split 70/30 et la validation croisée, suggérant une possible **tendance au surapprentissage**.

3. Overfitting :

○ Analyser la différence entre la performance en entraînement et en test/validation.

○ Observer si certains modèles ont une variance importante entre les folds ou

entre train/test.