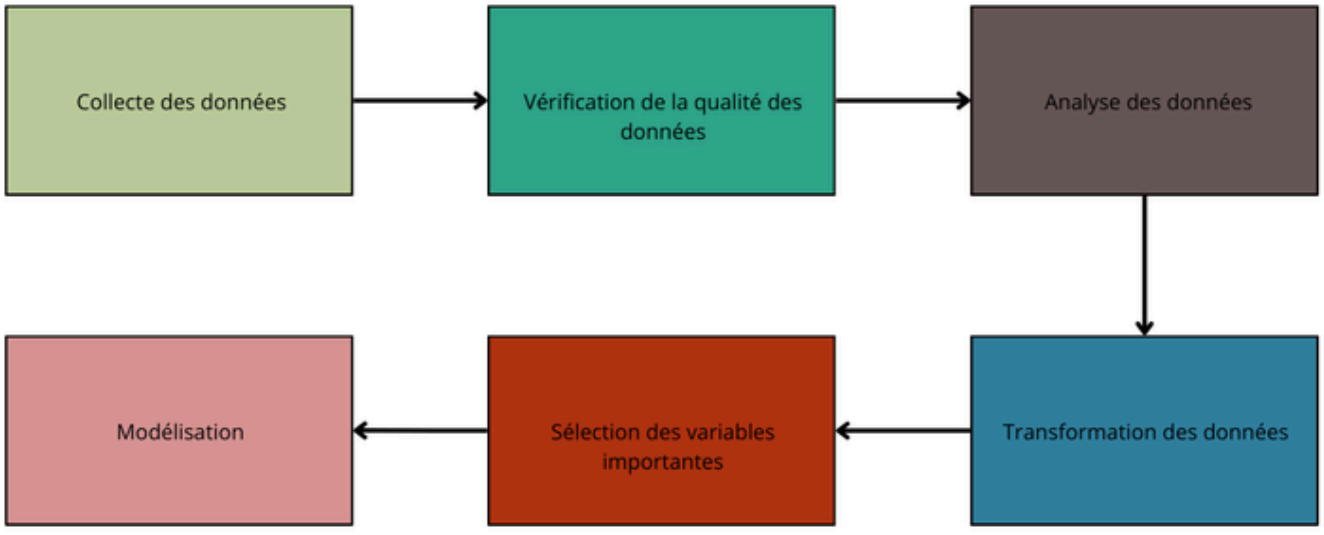
Rapport de projet

- Abstract

Dans ce travail, nous avons développé un modèle de classification en machine learning afin de prédire si un client souscrira à une offre bancaire. Pour ce faire, nous avons exploité un ensemble de données issues de campagnes de marketing direct menées par une institution bancaire portugaise entre 2008 et 2013. Ces données incluent des informations personnelles et financières des clients, des caractéristiques des campagnes de prospection, ainsi que des indicateurs socioéconomiques du contexte de l'époque.

Nous avons expérimenté deux algorithmes d'apprentissage supervisé : **les k-plus proches voisins (KNN) et les arbres de décision**. L'objectif était d'évaluer leurs performances respectives en termes de précision et de pertinence pour cette tâche de classification. L'évaluation des modèles a été réalisée à l'aide de métriques standards telles que l'exactitude, la précision, le rappel et le score F1. Les résultats obtenus permettent d'identifier le modèle le plus adapté à cette problématique et d'en analyser les performances selon les différents attributs des données.

Nous avons suivi la pipeline suivante:



- Collecte des données

Nous avons chargé les données à partir d'un fichier CSV dans un **DataFrame Pandas**. Ce DataFrame contient **41 188 lignes** et **21 colonnes**, incluant notre **variable cible (y)**, qui indique si un client a souscrit ou non à l'offre bancaire. Cette étape est essentielle pour explorer et préparer les données avant d'appliquer les modèles de machine learning.

- Vérification de la qualité des données

Certaines colonnes du dataset comportaient des valeurs "unknown", indiquant des données manquantes. Nous avons décidé de les considérer comme des **catégories à part entière** plutôt que de les supprimer ou de les remplacer, afin d'éviter d'introduire un biais dans notre analyse.

Les variables **job**, **marital**, **housing**, **loan** et **education** contiennent un faible pourcentage de "unknown" ($\lesssim 5\%$) et ont donc été conservées comme des modalités normales.

La variable **default**, en revanche, présente plus de **20% de valeurs "unknown"**, ce qui est significatif et nécessite une attention particulière dans l'analyse.

Nous avons procédé à la vérification des valeurs aberrantes dans les données. Après examen, nous avons constaté que celles identifiées n'étaient pas des erreurs ou des données incorrectes, mais plutôt des éléments inhabituels. Ces données ne semblaient pas correspondre aux tendances générales, mais elles ne constituaient pas nécessairement des anomalies ou des erreurs flagrantes.

- Analyse des données

En procédant à une analyse détaillée des données, nous avons observé que près de 90 % des personnes ont choisi de ne pas souscrire à l'offre qui leur a été proposée. Cela montre une forte réticence générale envers la proposition. Cependant, lorsque nous examinons les proportions de personnes ayant effectivement accepté de souscrire à l'offre, nous remarquons que la répartition des réponses positives en fonction de l'âge présente une forme assez normale, avec un taux de souscription particulièrement plus élevé chez les individus âgés de 29 à 50 ans. Ce groupe semble donc être celui qui réagit le plus favorablement à l'offre.

Par ailleurs, un autre aspect intéressant qui ressort de notre analyse est que les personnes ayant déjà été contactées lors d'une campagne précédente sont généralement plus enclines à refuser la souscription. Ce phénomène pourrait être dû à un effet de saturation ou à une certaine méfiance envers les campagnes répétées.

En outre, il est important de souligner que certaines catégories professionnelles montrent un taux de souscription plus élevé que d'autres. Cela pourrait être lié à des facteurs spécifiques à chaque secteur d'activité, comme la nature des emplois, les attentes ou les besoins spécifiques des individus. Enfin, nous avons également remarqué que l'efficacité de la campagne varie en fonction des mois de l'année, avec des périodes où le taux de souscription est plus important. Cela pourrait être lié à des facteurs saisonniers ou à des comportements de consommation spécifiques à certaines périodes.

- Transformation des données

Certaines variables de notre base de données présentent des distributions très déséquilibrées. Afin de remédier à ce problème, nous avons choisi d'ajouter de nouvelles colonnes dans notre jeu de données, contenant le logarithme des valeurs des variables à distribution déséquilibrée. L'objectif de cette transformation est de réduire l'impact des valeurs extrêmes et de rendre la distribution des données plus proche d'une distribution normale.

Une fois cette étape réalisée, nous avons procédé à une opération de **one-hot encoding** sur nos variables catégorielles. Cela a été nécessaire, car les modèles que nous comptons utiliser, tels que le KNN et l'arbre de décision, fonctionnent de manière plus optimale lorsque toutes les variables sont numériques. En effet, pour le KNN, l’utilisation de variables numériques est une exigence dans la bibliothèque scikit-learn.

Enfin, nous avons normalisé nos données pour nous assurer que toutes les variables soient sur la même échelle. Cette normalisation permet d'éviter que certaines variables, ayant des échelles très différentes, n'influencent de manière disproportionnée les résultats des modèles, notamment dans le cas du KNN, qui est particulièrement sensible aux différences d'échelle entre les variables.

- Sélection des variables importantes

En tenant compte de la taille de notre jeu de données après l'encodage, nous avons pris la décision d'analyser l'importance de certaines colonnes. Nous avons d'abord évalué la corrélation de chaque colonne avec notre variable cible. Les colonnes présentant une faible corrélation ont été identifiées et supprimées, car elles n'apportaient pas suffisamment d'information pour prédire la variable cible de manière significative.

Nous avons également détecté les colonnes ayant une très forte corrélation entre elles, ce qui pouvait indiquer une redondance dans les informations fournies. Ces colonnes, qui mesuraient en grande partie les mêmes relations avec la cible, ont été éliminées pour éviter le problème de multicolinéarité et alléger notre modèle.

Après avoir supprimé ces colonnes inutiles ou redondantes, nous avons obtenu un jeu de données de taille plus raisonnable, mieux adapté à l'entraînement des modèles. Cette réduction de la dimensionnalité permet également d'améliorer les performances du modèle tout en rendant l'entraînement plus rapide et plus efficace.

- Modélisation

Pour modéliser notre problème de classification, nous avons choisi d'utiliser deux algorithmes : un arbre de décision et le KNN (K-Nearest Neighbors). Étant donné que nos données sont fortement déséquilibrées, nous avons appliqué la méthode de suréchantillonnage SMOTE (Synthetic Minority Over-sampling Technique) pour équilibrer les classes et améliorer la performance de nos modèles.

Nous avons ensuite divisé notre jeu de données en deux ensembles : 80 % pour l'entraînement et 20 % pour les tests. Les résultats des tests se sont révélés très positifs, avec des scores de précision dépassant les 90 %, ce qui montre que nos modèles arrivent bien à prédire la variable cible.

Concernant l'évaluation de l'efficacité du KNN, nous avons exploré différents choix de la valeur de *k*. Nous avons constaté que lorsque la valeur de k est trop petit, cela peut affecter la precision de nos precisions. Lorsque k s'approche de 10, l'impact semble se stabiliser et la precision est optimale. Le graphe nous montre d'une valeur de K >= 5 donne des predictions acceptable.

Pour renforcer la performance de nos modèles, nous avons également intégré un modèle Ramdom Forest. Après avoir évalué ce modèle, nous avons observé que ses résultats étaient à peu près similaires à ceux obtenus avec les arbres de décision et le KNN, ce qui confirme que les approches basées sur ces algorithmes sont relativement performantes dans notre contexte, même après l'ajout de la Random Forest.