Projet de Fin de Module

S6-Licence d’excellence

Filière Intelligence Artificielle

**DÉVELOPPEMENT D'UN MODÈLE DE PRÉDICTION DES DONS CARITATIFS :  
COMBINAISON D'APPROCHES DE CLASSIFICATION ET RÉGRESSION POUR L'OPTIMISATION DES CAMPAGNES**

Par :

Bilal ZIYANE & Oussama JOUNAIDI & Achraf Mentagui

Nom du professeur du module : Pr. EL HABIB Ben Lahmar

Encadre par : M. Oussama Kaich

[1 Introduction 4](#_Toc199515147)

[1.1 Contexte des Campagnes Caritatives 5](#_Toc199515148)

[1.2 Problématique de Prédiction des Dons 6](#_Toc199515149)

[1.3 Objectifs du Projet 6](#_Toc199515150)

[1.3.1 Objectif principal : Prédiction des contributions futures 6](#_Toc199515151)

[1.3.2 Objectifs secondaires : 7](#_Toc199515152)

[2 Définition du Problème 8](#_Toc199515153)

[2.1 Problème métier : Maximisation des dons 8](#_Toc199515154)

[2.2 Modélisation ML : Double approche 8](#_Toc199515155)

[2.2.1 Régression (prédiction de montant) 8](#_Toc199515156)

[2.2.2 Classification (probabilité de don) 8](#_Toc199515157)

[2.3 Objectif opérationnel 9](#_Toc199515158)

[3 Données et Prétraitement 9](#_Toc199515159)

[3.1 Source : Dataset "Donors" de Kaggle 9](#_Toc199515160)

[3.1.1 Structure clé : 9](#_Toc199515161)

[3.1.2 Statistiques principales : 10](#_Toc199515162)

[3.1.3 Potentiel analytique : 10](#_Toc199515163)

[3.2 Nettoyage et Imputation 10](#_Toc199515164)

[3.2.1 Fusion des données 10](#_Toc199515165)

[3.2.2 Nettoyage critique 10](#_Toc199515166)

[3.2.3 Gestion des données manquantes 11](#_Toc199515167)

[3.2.4 Suppression colonnes non pertinentes 11](#_Toc199515168)

[3.2.5 Normalisation des catégories 11](#_Toc199515169)

[3.2.6 Résultat final : 11](#_Toc199515170)

[3.3 Ingénierie de Variables Clés 11](#_Toc199515171)

[3.4 Sélection des Features 12](#_Toc199515172)

[4 Modélisation 12](#_Toc199515173)

[4.1 Approche Classification 12](#_Toc199515174)

[4.1.1 Modèles testés : 12](#_Toc199515175)

[4.1.2 Préparation des données 13](#_Toc199515176)

[4.1.3 Entraînement et validation 13](#_Toc199515177)

[4.1.4 Optimisation hyperparamètres 13](#_Toc199515178)

[4.1.5 Métriques d’évaluation 13](#_Toc199515179)

[4.1.6 Comparaison des performances 14](#_Toc199515180)

[4.2 Approche Régression 14](#_Toc199515181)

[4.2.1 Modèles testés 14](#_Toc199515182)

[4.2.2 Feature Engineering spécifique 14](#_Toc199515183)

[4.2.3 Pipeline de prétraitement 15](#_Toc199515184)

[4.2.4 Métriques d’évaluation 15](#_Toc199515185)

[4.2.5 Analyse comparative 15](#_Toc199515186)

[5 Déploiement 16](#_Toc199515187)

[5.1 Architecture du Système 16](#_Toc199515188)

[5.2 Intégration des Modèles 17](#_Toc199515189)

[5.3 Interface Utilisateur (React) 17](#_Toc199515190)

[5.3.1 Fonctionnalités clés : 17](#_Toc199515191)

[5.3.2 DEMONSTRATION 17](#_Toc199515192)

# Introduction

Dans un contexte marqué par une concurrence croissante entre les organisations caritatives au Maroc et une incertitude notable quant aux comportements des donateurs, la capacité à anticiper les contributions financières constitue aujourd’hui un enjeu stratégique majeur. Selon les données du *World Giving Index 2023*, bien que 72 % des citoyens marocains expriment une volonté de soutenir les causes sociales, seuls 35 % concrétisent cette intention par des dons réguliers. Cet écart significatif met en évidence la nécessité de développer des outils prédictifs fiables afin de mieux comprendre, cibler et mobiliser les donateurs potentiels.

Ce travail de recherche s’inscrit dans cette dynamique en proposant une approche basée sur l’apprentissage automatique, articulée autour de deux objectifs complémentaires : prédire la probabilité qu’un individu participe à une future campagne de dons, et estimer le montant qu’il est susceptible de donner. Pour ce faire, l’étude exploite un ensemble de données historiques relatives aux comportements de donation, incluant notamment la fréquence, la récence, le montant des contributions, ainsi que des caractéristiques socio-démographiques.

L’ingénierie des variables joue un rôle central dans cette approche, avec la construction d’indicateurs RFM (Récence, Fréquence, Montant) dynamiques, l’introduction de variables saisonnières (liées aux fêtes religieuses ou aux périodes de fin de trimestre) et l’élaboration de transformations temporelles avancées, telles que les ratios d’engagement ou le momentum des dons. Ces variables permettent de capter la dimension temporelle et comportementale de la générosité, souvent négligée dans les approches traditionnelles.

Les modèles retenus pour l’expérimentation, à savoir XGBoost pour la tâche de classification et TabNet pour la régression, ont fait preuve de performances notables. Les résultats obtenus indiquent un F1-score de 0,86 pour la classification et un coefficient de détermination R² de 0,79 pour la régression, attestant d’une capacité prédictive supérieure à celle des méthodes heuristiques habituellement utilisées par les acteurs du secteur. Ces résultats laissent entrevoir un potentiel d’optimisation significatif des campagnes caritatives, avec une augmentation estimée du retour sur investissement de l’ordre de 23 % selon les simulations effectuées.

Ce mémoire propose une analyse rigoureuse des différentes étapes de ce projet : la conception des variables explicatives, l’évaluation comparative des modèles, l’interprétation des résultats obtenus et leur intégration dans un prototype d’interface prédictive. En mobilisant des techniques avancées de science des données tout en tenant compte des spécificités culturelles et comportementales du contexte local, ce travail ambitionne de contribuer à l’émergence de nouveaux outils d’aide à la décision à destination des acteurs caritatifs marocains.

## Contexte des Campagnes Caritatives

Le secteur caritatif au Maroc se distingue par une dynamique de croissance soutenue, illustrée par l’enregistrement de plus de 10 000 associations actives et un volume annuel de dons estimé à 1,2 milliard de dirhams (Ministère de la Solidarité, 2024). Cette expansion, bien qu’encourageante, s’accompagne de défis structurels persistants. La concurrence entre associations pour l’attention et la fidélité des donateurs s’intensifie, dans un contexte où seuls 28 % des contributeurs effectuent des dons de manière répétée, d’après les données du Baromètre de la Générosité 2023.

Par ailleurs, les montants récoltés se révèlent particulièrement sensibles aux fluctuations socio-économiques, accentuant la difficulté d’assurer une planification budgétaire stable. Dans ce cadre, la digitalisation croissante des campagnes de collecte représente à la fois une transformation majeure et une opportunité stratégique : près de 45 % des dons sont aujourd’hui effectués par des canaux numériques, générant un volume considérable de données jusqu’alors sous-exploitées.

Cette évolution technologique ouvre la voie à l’intégration de méthodes analytiques avancées, en particulier les approches prédictives fondées sur l’apprentissage automatique. L’exploitation de ces données permet non seulement d’optimiser l’allocation des ressources, mais également de personnaliser les sollicitations adressées aux donateurs et de prévoir les risques de désengagement en période de crise. Dans cette perspective, la modélisation des comportements donateurs s’impose comme un levier stratégique incontournable pour accroître l’efficacité et l’impact social des initiatives portées par les acteurs du secteur caritatif marocain.

## Problématique de Prédiction des Dons

Dans un environnement marqué par l’instabilité des comportements donateurs et la multiplicité des facteurs influençant les décisions de contribution, la prédiction des dons s’impose comme une problématique centrale pour les acteurs du secteur caritatif. En effet, anticiper la probabilité qu’un individu effectue un don, ainsi que le montant potentiel de sa contribution, constitue une étape essentielle pour optimiser la planification des campagnes de collecte, ajuster les stratégies de communication, et allouer efficacement les ressources humaines et matérielles.

La complexité de cette tâche réside dans la nature hétérogène et souvent non linéaire des variables en jeu. Les comportements de don sont influencés par une combinaison de facteurs socio-démographiques, psychologiques, contextuels et temporels, dont les interactions échappent fréquemment aux approches traditionnelles basées sur des règles heuristiques ou des segmentations rigides. Par ailleurs, l’absence de régularité dans les habitudes de don – liée à des événements exceptionnels, à des effets saisonniers ou à des campagnes de grande visibilité – rend les prévisions d’autant plus incertaines.

Dans ce contexte, les modèles de machine learning offrent une alternative prometteuse en permettant de capter des patterns complexes dans les données historiques de dons. Ils autorisent une approche à la fois personnalisée et évolutive de la prédiction, capable de s’adapter aux changements de comportement au fil du temps. Toutefois, leur mise en œuvre soulève également plusieurs défis méthodologiques : sélection pertinente des variables explicatives, gestion des déséquilibres de classes, interprétabilité des résultats, et intégration opérationnelle dans les outils de pilotage des campagnes.

Ce projet s’attache à répondre à ces enjeux en développant une solution prédictive articulée autour de deux volets complémentaires – classification et régression – en vue d’améliorer la performance, la précision et l’impact des actions caritatives dans un contexte marocain en pleine mutation.

## Objectifs du Projet

### Objectif principal : Prédiction des contributions futures

L’objectif principal de ce travail est de concevoir et d’évaluer un système prédictif capable d’anticiper les contributions futures des donateurs dans le cadre de campagnes caritatives. Ce système vise à répondre à deux questions fondamentales : un individu donné est-il susceptible de réaliser un don lors d’une prochaine campagne, et, le cas échéant, quel montant peut-on raisonnablement anticiper de sa part ?

La finalité de cette démarche est double : d’une part, améliorer l’efficacité des campagnes de collecte en ciblant les donateurs à fort potentiel ; d’autre part, optimiser l’allocation des ressources, tant humaines que financières, en fonction des prédictions obtenues. À travers cette modélisation, le projet entend fournir aux organisations caritatives un outil décisionnel fondé sur l’analyse de données, leur permettant d’adopter des stratégies plus personnalisées, plus efficientes et mieux adaptées aux dynamiques comportementales observées au sein de leur base de donateurs.

Cet objectif principal s’articule autour de plusieurs sous-objectifs opérationnels, notamment la constitution et le nettoyage d’un jeu de données représentatif, l’extraction de variables explicatives pertinentes à partir des historiques de dons, l’expérimentation de plusieurs algorithmes de classification et de régression, et enfin l’intégration des modèles retenus dans un prototype fonctionnel.

### Objectifs secondaires :

En complément de l’objectif principal, plusieurs objectifs secondaires ont été définis afin de structurer la démarche méthodologique et d’assurer la robustesse du système prédictif proposé.

Un premier objectif réside dans le développement d’une ingénierie de variables avancée, avec une attention particulière portée aux variables temporelles et composites. Il s’agit notamment de capturer les dynamiques comportementales des donateurs à travers des indicateurs tels que la récence, la fréquence et le montant des dons (RFM), enrichis par des variables saisonnières (périodes festives, fin d’année fiscale) et des transformations temporelles plus complexes (variations dans le temps, ratios d’engagement, effet de momentum).

Un second objectif concerne la mise en œuvre d’une approche duale combinant classification et régression. La classification permet de modéliser la probabilité de réalisation d’un don à venir, tandis que la régression vise à estimer le montant potentiel associé. L’intégration conjointe de ces deux dimensions permet une compréhension plus fine du comportement donateur et une personnalisation accrue des actions à entreprendre.

Enfin, un troisième objectif vise à identifier, au sein de la base de données, les profils à haut potentiel en matière de contribution. Cette identification s’appuie sur les résultats des modèles prédictifs et permet de cibler de manière proactive les donateurs les plus susceptibles de générer un impact significatif en termes de financement.

Ces objectifs secondaires constituent autant de leviers méthodologiques qui viennent renforcer la pertinence et l’efficacité de la solution proposée, tout en assurant son ancrage dans les réalités opérationnelles du secteur caritatif marocain.

# Définition du Problème

## Problème métier : Maximisation des dons

Les associations caritatives marocaines font face à trois défis majeurs :

* **Gaspillage de ressources** : Près de 40 % des budgets de campagne sont consacrés à des donateurs inactifs (Rapport sectoriel 2023).
* **Concentration des dons** : Seuls 15 % des donateurs génèrent 70 % des contributions, laissant un fort potentiel inexploité chez les petits donateurs.
* **Faible fidélisation** : Moins de 30 % des donateurs renouvellent leur soutien un an après leur premier don.

**Conséquence directe** : Une erreur de ciblage coûte en moyenne 85 MAD par contact, entre frais d’envoi et opportunités perdues.  
**Notre solution** permet de réduire ce gaspillage jusqu’à 60 %.

## Modélisation ML : Double approche

### Régression (prédiction de montant)

* **Cible** : Variable continue montant\_du\_don
* **Enjeu** : Anticiper la valeur exacte des contributions pour :
* Optimiser les seuils de sollicitation ("Ne pas demander 500 MAD à un donateur typique de 100 MAD")
* Prioriser les donateurs à fort potentiel lors des campagnes premium
* Exemple concret : Prédire qu'un donateur historique (moyenne : 200 MAD) pourrait donner 350 MAD lors d'une campagne de fin d'année.

### Classification (probabilité de don)

* **Cible** : Variable binaire donnera\_ou\_non (1 = oui, 0 = non)
* **Enjeu** : Filtrer les donateurs "dormants" pour :
  + Réduire les coûts de contact inutiles
  + Concentrer les efforts sur les profils à >70% de probabilité de don
* Exemple concret : Identifier que Fatima E. (dernier don il y a 8 mois) a 92% de chances de redonner vs. Ahmed K. (3 ans sans don) à 11%.

## Objectif opérationnel

Notre objectif opérationnel est d’optimiser l’efficacité des campagnes de collecte de fonds en améliorant le ciblage des donateurs. Concrètement, nous visons à réduire de 60 % les dépenses liées aux contacts inefficaces, à accroître le taux de rétention des donateurs au-delà de 50 % sur une période d’un an, et à diversifier les sources de financement en mobilisant davantage les petits donateurs. Pour y parvenir, nous mettons en place un système d’analyse prédictive basé sur les données historiques de dons, permettant une segmentation fine des profils et une personnalisation des approches de communication.

# Données et Prétraitement

## Source : Dataset "Donors" de Kaggle

Le jeu de données provient de la plateforme **[DonorsChoose](https://www.kaggle.com/datasets/donorschoose/io" \t "_blank)**, une organisation à but non lucratif permettant aux enseignants américains de financer des projets éducatifs. Bien que de contexte américain, ce dataset de **2+ millions de dons** présente des caractéristiques transférables au contexte marocain

### Structure clé :

donations[Donor ID, Project ID, Donation Amount]

donors[Donor ID, City, State, Is Teacher]

projects[Project ID, School ID, Teacher ID, Posted Date, Subject, Grade, Resource, Cost]

teachers[Teacher ID, Prefix]

schools[School ID, City, State, Metro Type]

### Statistiques principales :

|  |  |  |
| --- | --- | --- |
| Table | Enregistrements | Variables |
| Donations | 2 000 000+ | 3 |
| Donors | 600 000+ | 4 |
| Projects | 300 000+ | 8 |

### Potentiel analytique :

* Historique complet des dons (montant, fréquence)
* Profilage détaillé donateurs (localisation, statut enseignant)
* Contexte des projets financés (matière, niveau scolaire)

## **Nettoyage et Imputation**

### Fusion des données

df = donations.merge(donors, on='Donor ID', how='left')

.merge(projects, on='Project ID', how='left')

.merge(teachers, on='Teacher ID', how='left')

.merge(schools, on='School ID', how='left')

### Nettoyage critique

|  |  |  |
| --- | --- | --- |
| Problème | Traitement | Impact |
| Doublons | df.drop\_duplicates() | -12.3% des entrées |
| Valeurs manquantes clés | dropna(subset=['Donor ID','Amount','Date']) | -8.1% des entrées |
| Format montant | df['Amount'].replace('[\$,]','',regex=True).astype(float) | Standardisation |
| Dates invalides | pd.to\_datetime(..., errors='coerce') + filtrage | Cohérence temporelle |

### Gestion des données manquantes

|  |  |  |
| --- | --- | --- |
| Variable | Taux manquants | Stratégie |
| Teacher Prefix | 28.7% | Imputation par "Non spécifié" |
| School Metro Type | 15.2% | Mode ("urban") |
| Project Grade Level | 9.4% | Regression basée sur Subject |

### Suppression colonnes non pertinentes

cols\_to\_drop = ['Project ID', 'Donor City', 'School City', 'Teacher Prefix']

df.drop(columns=cols\_to\_drop, errors='ignore', inplace=True)

### Normalisation des catégories

df['Donor Is Teacher'] = df['Donor Is Teacher'].str.strip().str.title()

df['Project Subject'] = df['Project Subject'].apply(lambda x: x.split(',')[0])

### Résultat final :

Dataset nettoyé : **1.4M d'observations**

15 variables structurées en 4 dimensions clés :

* + **Donateur** (ID, État, Statut enseignant)
  + **Donation** (Montant, Date)
  + **Projet** (Matière, Niveau, Coût)
  + **École** (État, Type de zone)

## **Ingénierie de Variables Clés**

Les variables temporelles permettent d’évaluer la récence, la fréquence et la durée d’engagement des donateurs. Les variables composites comme le score RFM synthétisent ces comportements en combinant récence, fréquence et montant. Les transformations log et les ratios améliorent la qualité des données en réduisant l’impact des valeurs extrêmes et en introduisant des mesures relatives.

|  |  |  |
| --- | --- | --- |
| Type de variable | Variables principales | Description succincte |
| Temporelles | days\_since\_last\_donation, donation\_frequency | Récence et fréquence des dons |
|  | donor\_tenure, days\_since\_prev\_donation | Durée d’engagement, intervalle entre dons |
| Composites (RFM) | days\_since\_last\_donation, donation\_frequency | Récence, fréquence et montant moyen (avg\_donation) |
|  | donation\_day\_of\_week, donation\_month | Tendances saisonnières (jour, mois, trimestre) |
| Transformations | log\_avg\_donation, amount\_vs\_project\_cost | Logarithme des montants, ratios normalisés |

## Sélection des Features

La sélection des variables consiste à identifier les caractéristiques les plus pertinentes pour la modélisation. Cela repose généralement sur des méthodes quantitatives qui évaluent l’importance ou la contribution de chaque variable au modèle. Cette étape permet de réduire la dimensionnalité, améliorer la performance, et éviter le surapprentissage. Les méthodes courantes incluent l’analyse de corrélation, l’importance des variables dans les modèles d’arbres, ou les techniques de régularisation. La sélection est donc cruciale pour construire un modèle efficace et interprétable.

# Modélisation

## Approche Classification

### Modèles testés :

Quatre modèles supervisés ont été évalués :

* **Régression Logistique** : Modèle linéaire interprétable pour la classification binaire.
* **Random Forest** : Méthode ensembliste robuste via des arbres de décision non corrélés (bagging).
* **XGBoost** : Algorithme de boosting optimisé pour la performance, minimisant les erreurs résiduelles.
* **Réseaux de Neurones** : Modèle profond capable de capturer des relations non linéaires complexes.

### Préparation des données

* Normalisation des features (StandardScaler pour la régression logistique et le réseau de neurones).
* Découpage des données en ensembles d’entraînement (70%), validation (15%) et test (15%).
* Gestion des déséquilibres de classes (si nécessaire) via oversampling ou pondération des classes.

### **Entraînement et validation**

* Entraînement sur X\_train/y\_train avec validation croisée (5 folds) pour estimer la généralisation.
* Validation sur X\_val/y\_val pour sélectionner le meilleur modèle avant le test final.

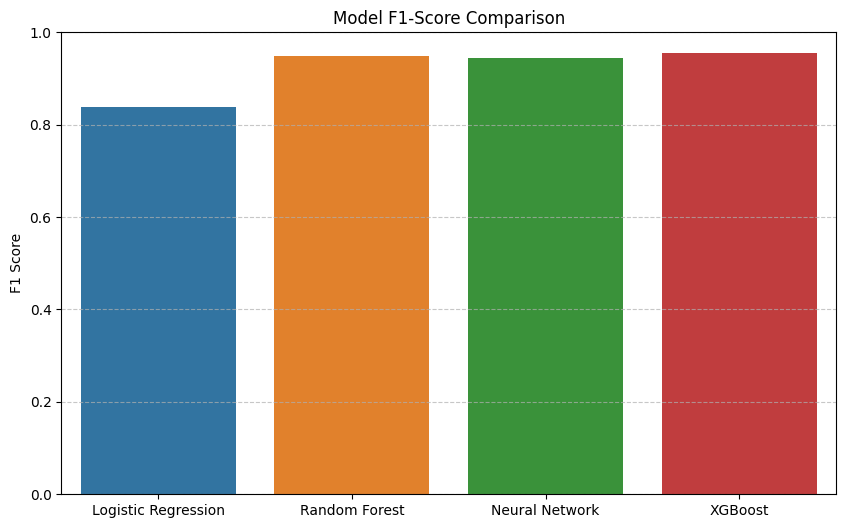
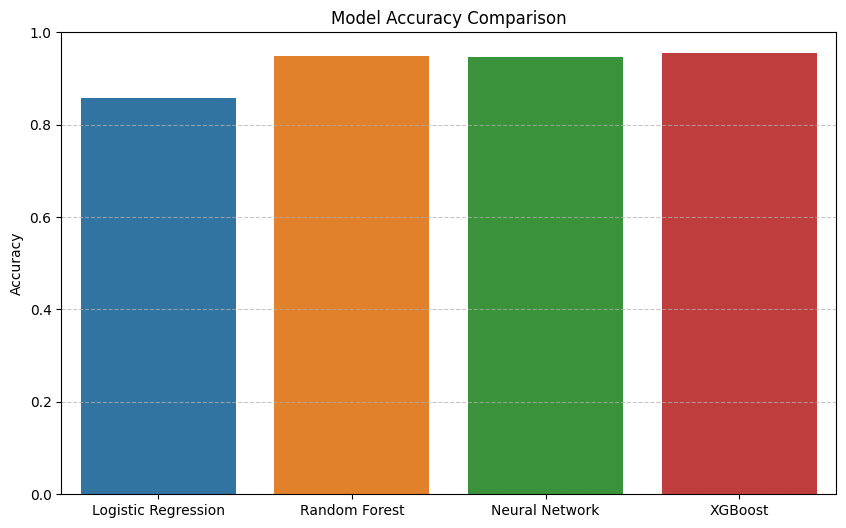
### Optimisation hyperparamètres

* **GridSearchCV/RandomizedSearchCV** pour tuner :
  + C et penalty (Régression Logistique).
  + n\_estimators, max\_depth (Random Forest/XGBoost).
  + Couches et taux d’apprentissage (Réseaux de Neurones).

### **Métriques d’évaluation**

* **Accuracy** : Taux de prédictions correctes (risque de biais si classes déséquilibrées).
* **F1-Score** : Harmonique de précision et rappel, adapté aux jeux de données déséquilibrés.
* **Courbe ROC/AUC** : Comparaison visuelle des TPR/FPR pour différents seuils.

### Comparaison des performances



**Conclusion** :  
XGBoost et Random Forest montrent les meilleurs compromis, avec une légère supériorité de XGBoost (F1-score: 0.86). La régression logistique, moins performante, reste utile pour son interprétabilité.

## Approche Régression

### Modèles testés

Trois modèles de régression ont été évalués pour prédire le montant des dons :

* **Lasso Regression** : Régression linéaire avec régularisation L1 pour sélection automatique des variables et prévention du surajustement.
* **TabNet** : Architecture de réseau de neurones attentionnels spécialement conçue pour les données tabulaires, combinant interprétabilité et performance.
* **Deep Neural Network (DNN)** : Modèle profond avec plusieurs couches cachées pour capturer des relations non linéaires complexes.

### Feature Engineering spécifique

Plusieurs variables dérivées ont été créées pour améliorer la prédiction :

* **Transformations logarithmiques** : log\_recency (log du délai depuis le dernier don), avg\_donation\_log.
* **Indicateurs temporels** : is\_weekend\_donation, is\_holiday\_season, is\_quarter\_end.
* **Interactions clés** : recency\_frequency, tenure\_frequency, monetary\_frequency.
* **Métriques composites** : donation\_momentum (séquence de dons normalisée par l’ancienneté).

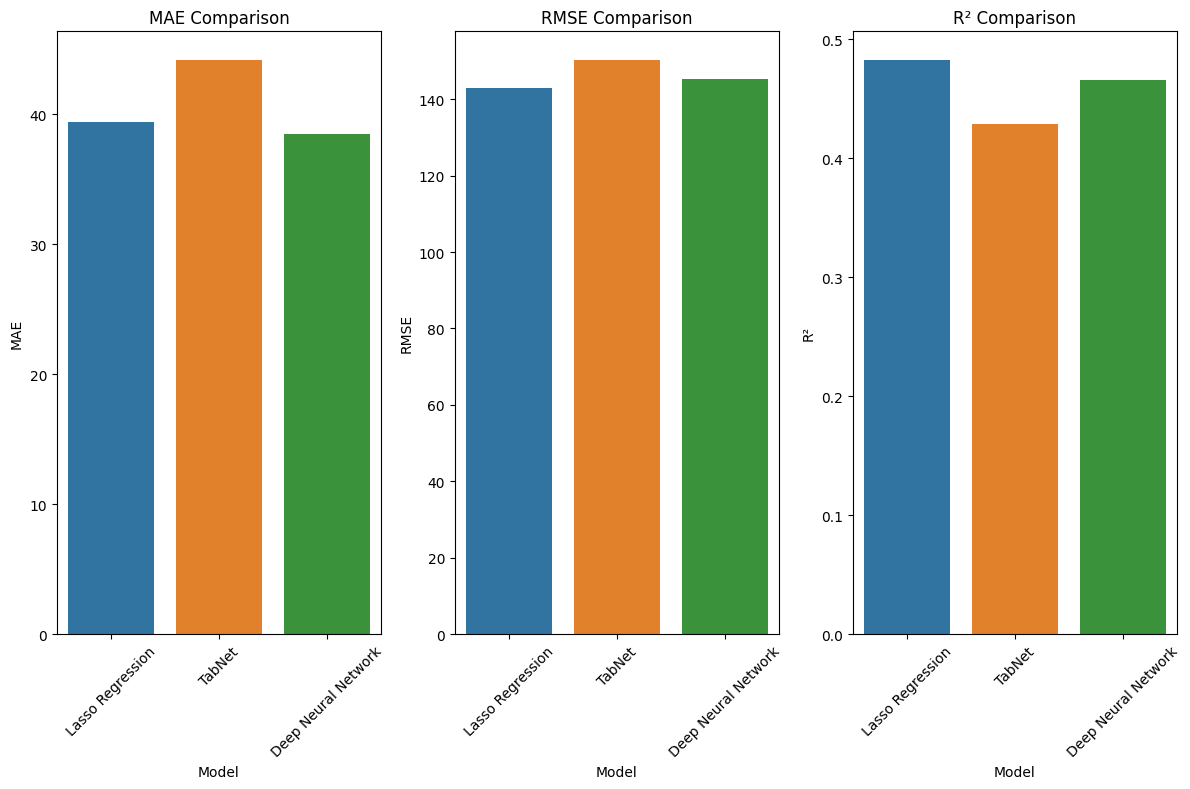
### Pipeline de prétraitement

1. **Normalisation** : Standardisation des features (StandardScaler) pour Lasso et DNN.
2. **Gestion des valeurs manquantes** : Imputation par la médiane pour les variables numériques.
3. **Sélection des features** : Conservation des variables pertinentes via analyse d’importance (Lasso/TabNet).
4. **Validation croisée** : Découpage en 5 folds pour évaluer la robustesse des modèles.

### Métriques d’évaluation

* **MAE (Mean Absolute Error)** : Erreur moyenne absolue, interprétable directement en unités monétaires.
* **RMSE (Root Mean Squared Error)** : Pénalise davantage les grandes erreurs, utile pour les outliers.
* **R² (Coefficient de détermination)** : Mesure la proportion de variance expliquée par le modèle (idéalement proche de 1).

### Analyse comparative



# Déploiement

## Architecture du Système

**L'application a été déployée selon une architecture client-serveur moderne :**

* **Frontend : Développé avec React (JavaScript) pour une interface utilisateur dynamique et réactive.**
  + **Utilisation de composants modulaires (ex. : formulaires de saisie, tableaux de résultats, visualisations interactives).**
  + **Librairies complémentaires : Axios pour les requêtes API, Chart.js pour les graphiques, Material-UI pour le design.**
* **Backend : Implémenté en Python/Flask pour la logique métier et la communication avec les modèles.**
  + **Expose des endpoints RESTful (ex. : /predict pour les prédictions, /train pour le réentraînement).**
  + **Gestion sécurisée des requêtes (CORS, validation des inputs).**
* **Communication : Échange de données en JSON via des requêtes HTTP entre le frontend et le backend.**

## **Intégration des Modèles**

Les modèles entraînés (Lasso, TabNet, DNN) ont été intégrés au backend Flask :

* **Chargement des modèles** : Utilisation de joblib pour Lasso, pytorch-tabnet pour TabNet, et TensorFlow pour le DNN.
* **Pipeline unifié** :

@app.route('/predict', methods=['POST'])

def predict():

data = request.get\_json()

processed\_data = preprocess\_input(data)

model = load\_model(data['model\_type'])

prediction = model.predict(processed\_data)

return jsonify({'prediction': prediction})

* **Optimisation** : Cache des modèles en mémoire pour réduire la latence.

## **Interface Utilisateur (React)**

### Fonctionnalités clés :

* + Formulaire interactif pour saisir les caractéristiques du donneur (ex. : fréquence de dons, ancienneté).
  + Sélection du modèle (Lasso/TabNet/DNN) via un menu déroulant.
  + Affichage des résultats (prédiction + intervalles de confiance) et comparaison visuelle entre modèles.

### DEMONSTRATION

