

# **CLASSIFICATION DIABETE**

**Réalisé par :**

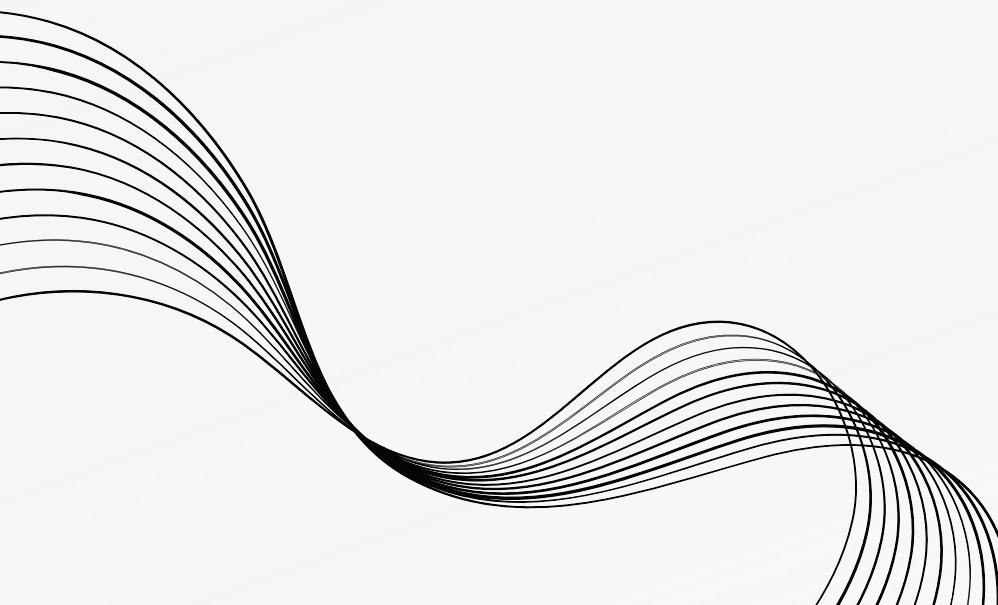
**Oussama TAZI**

**Houssam EL AZAMI EL IDRISI**

**Encadré par :**

**Mme. Rajaa Charifi**

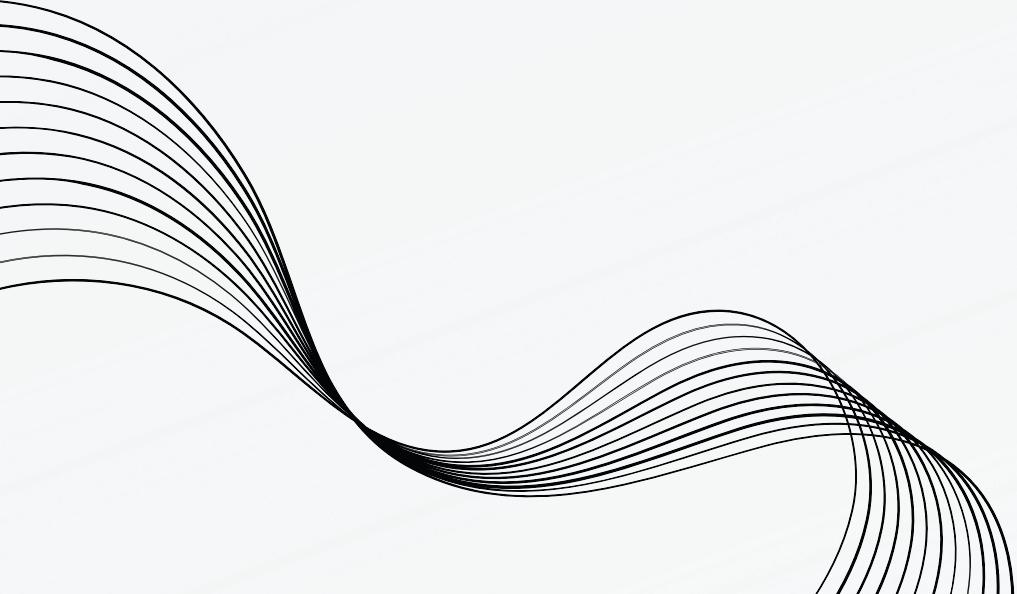
# INTRODUCTION



# PLAN

- 01** PROBLEMATIQUE
- 02** DESCRIPTION DU DATASET
- 03** PROCESSUS D'ANALYSE ET D'ENTRAINEMENT
- 04** CHOIX DU MODEL
- 05** ANALYSE DES RESULTATS OBTENUES

# **ÉNONCÉ DU PROBLÈME**



# PROBLEMATIQUE



- Le diabète est une maladie chronique en constante augmentation, avec environ 12,4% de la population au Maroc en étant atteinte. Le dépistage précoce est essentiel pour prévenir des complications graves. Cependant, la détection rapide et précise des individus à risque parmi une grande population demeure un défi majeur, en raison de la diversité des facteurs de risque, des symptômes et des comportements individuels. Le problème réside dans la difficulté d'identifier efficacement les personnes susceptibles de développer un diabète, retardant ainsi la prise en charge et aggravant les conséquences sanitaires.

# OBJECTIFS



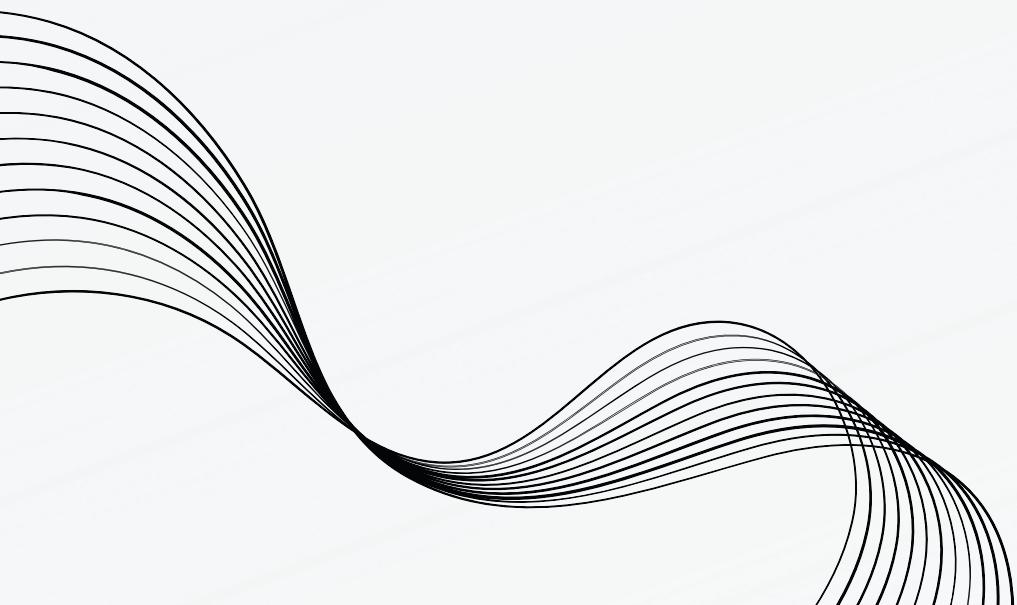
- Développer des modèles de classification du diabète.
- Comparer la performance des modèles (Régression Logistique, CatBoost, Random Forest).

# L'IMPORTANCE DE LA RÉSOLUTION DE CE PROBLÈME ?

- Réduire les risques de complications graves liées au diabète.
- Améliorer la gestion de la santé publique avec des dépistages ciblés.
- Optimiser les traitements et réduire les coûts médicaux à long terme.



# **DESCRIPTION DU DATASET**



# DATASET ET PREPROCESSING

## DETAILS DU DATASET

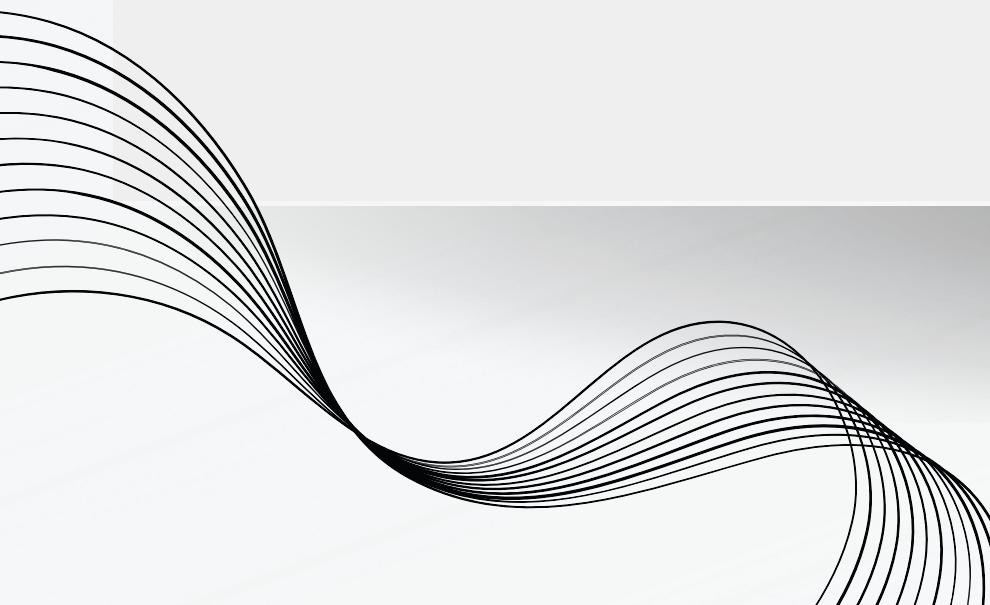
- Source : Kaggle - Diabetes prediction dataset
- Taille 3.81MB



# DATASET ET PREPROCESSING

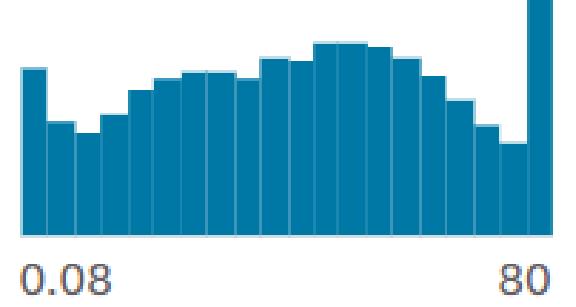
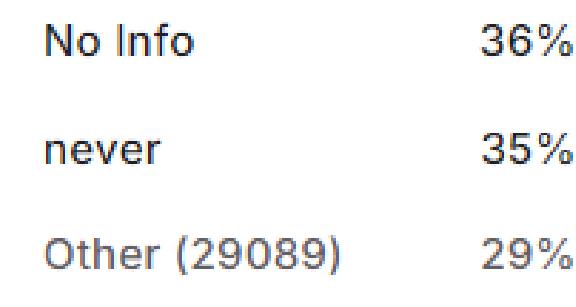
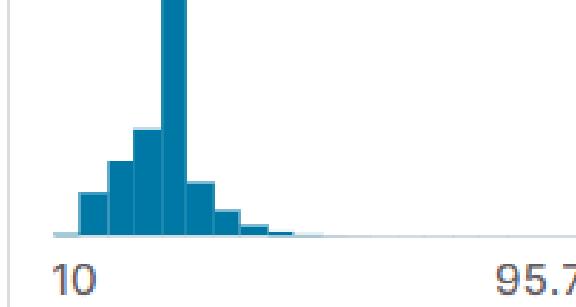
## CHOIX DU DATASET

- Contient des données de patients réelles, anonymisées pour protéger la confidentialité des individus.
- Offre plusieurs features pour un aspect global sur la santé générale de chaque patient.
- Pertinence pour la problématique.



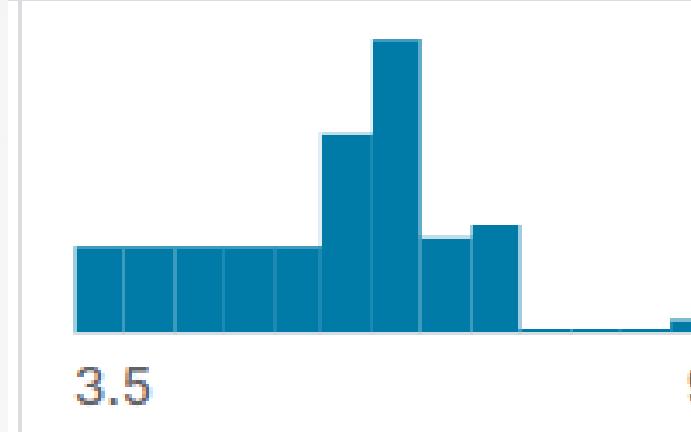
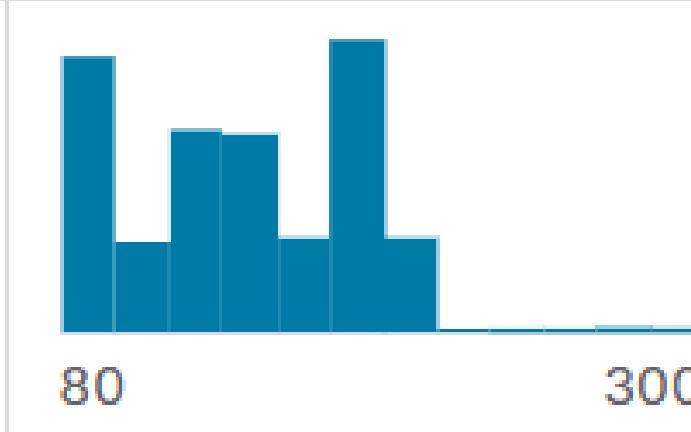
# DATASET ET PREPROCESSING

## STRUCTURE DE DATASET

# gender	Female Male Other (18)	59% 41% 0%	# age	Age is an important factor as diabetes is more commonly diagnosed in older adults. Age ranges from 0.08 to 80+.	# hypertension	Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. It	# heart_disease	Heart disease is another medical condition that is associated with an increased risk of developing diabetes. It	# smoking_history	Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated	# bmi	BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a
												

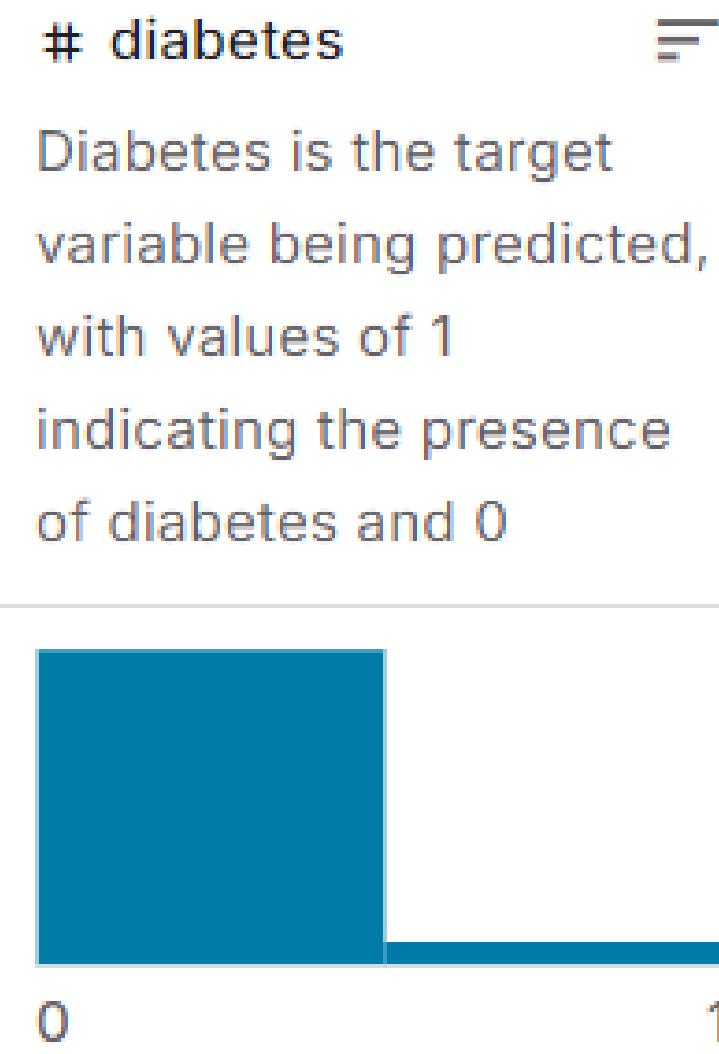
# DATASET ET PREPROCESSING

## STRUCTURE DE DATASET

# HbA1c_level	# blood_glucose_le...	# diabetes
HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher	Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose	Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0
		

# DATASET ET PREPROCESSING

## STRUCTURE DE DATASET



- Le Dataset n'est pas équilibré, ce qui engendra l'équilibration du dataset après.

# DATASET ET PREPROCESSING

## PREPROCESSING ET CLEANING

- Nous avons effectué le prétraitement nécessaire sur le dataset, qui consiste à traiter deux problèmes majeurs : les données redondantes et manquantes, en utilisant les bibliothèques Python :
  - pandas
  - sklearn.preprocessing.
- Lors du traitement des deux problèmes majeurs mentionnés ci-dessus, nous avons choisi de supprimer les lignes contenant des valeurs redondantes. Tandis que nous n'avons pas de valeurs manquantes



# DATASET ET PREPROCESSING

## FEATURE ENGINEERING

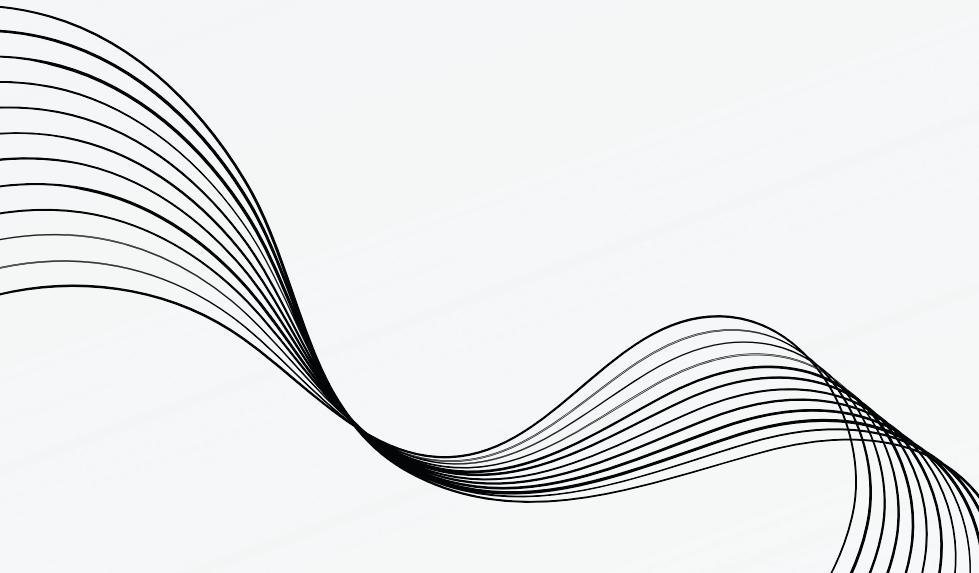
- Nous avons aussi fait le feature engineering pour obtenir des colonnes supplémentaires, permettant au modèle d'identifier toutes les patterns et relations potentielles, ce qui améliore sa capacité à prédire avec précision la présence de diabète.
- Les colonnes ajoutées incluent l'âge, l'IMC, les niveaux d'HbA1c et de glucose sanguin, ainsi que des variables liées au genre, à l'âge et aux seuils critiques pour améliorer la détection des patterns de diabète.



# **DATASET ET PREPROCESSING**

## **NORMALISATION DES DONNES**

- Pour la normalisation, nous avons utilisé StandardScaler, afin de standardiser les variables numériques et garantir que toutes les caractéristiques aient une échelle similaire, ce qui améliore la performance du modèle



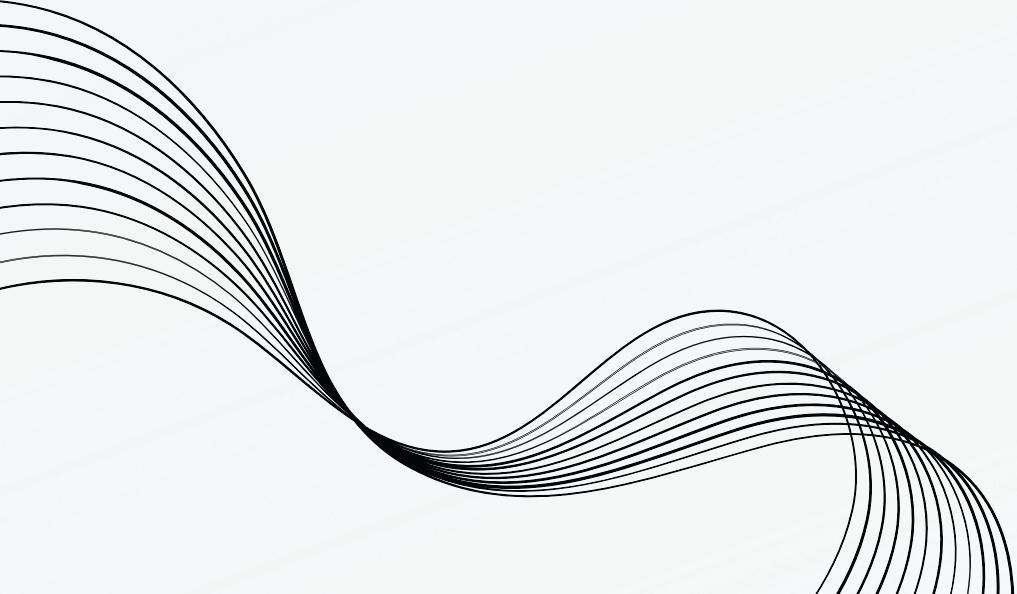
# **DATASET ET PREPROCESSING**

## **SPLITTING DU DATASET**

- Nous avons divisé notre dataset en quatre parties :
  - Training (70%)
  - Validation (10%)
  - Testing (20%)



# MÉTHODOLOGIE



# MODELES ETUDIÉS

Logitic Regression

CatBoost

Random Forest

# Métriques d'évaluation



## ACCURACY

La proportion de prédictions correctes sur l'ensemble des prédictions effectuées, mesurant l'exactitude globale du modèle.

## PRECISION

La proportion de prédictions positives correctes sur le total des prédictions positives, indiquant la fiabilité des prédictions positives.

## F1-SCORE

La moyenne harmonique de la précision et du rappel, fournissant un équilibre entre les deux métriques.

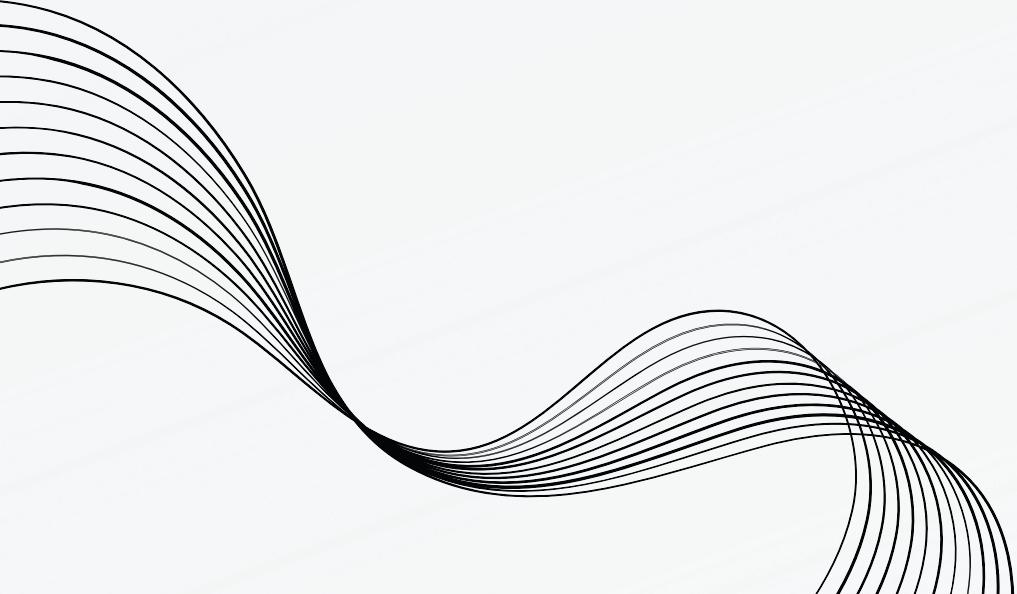
## ROC AUC

La surface sous la courbe ROC, mesurant la capacité du modèle à distinguer entre les classes positives et négatives sur différents seuils de classification.

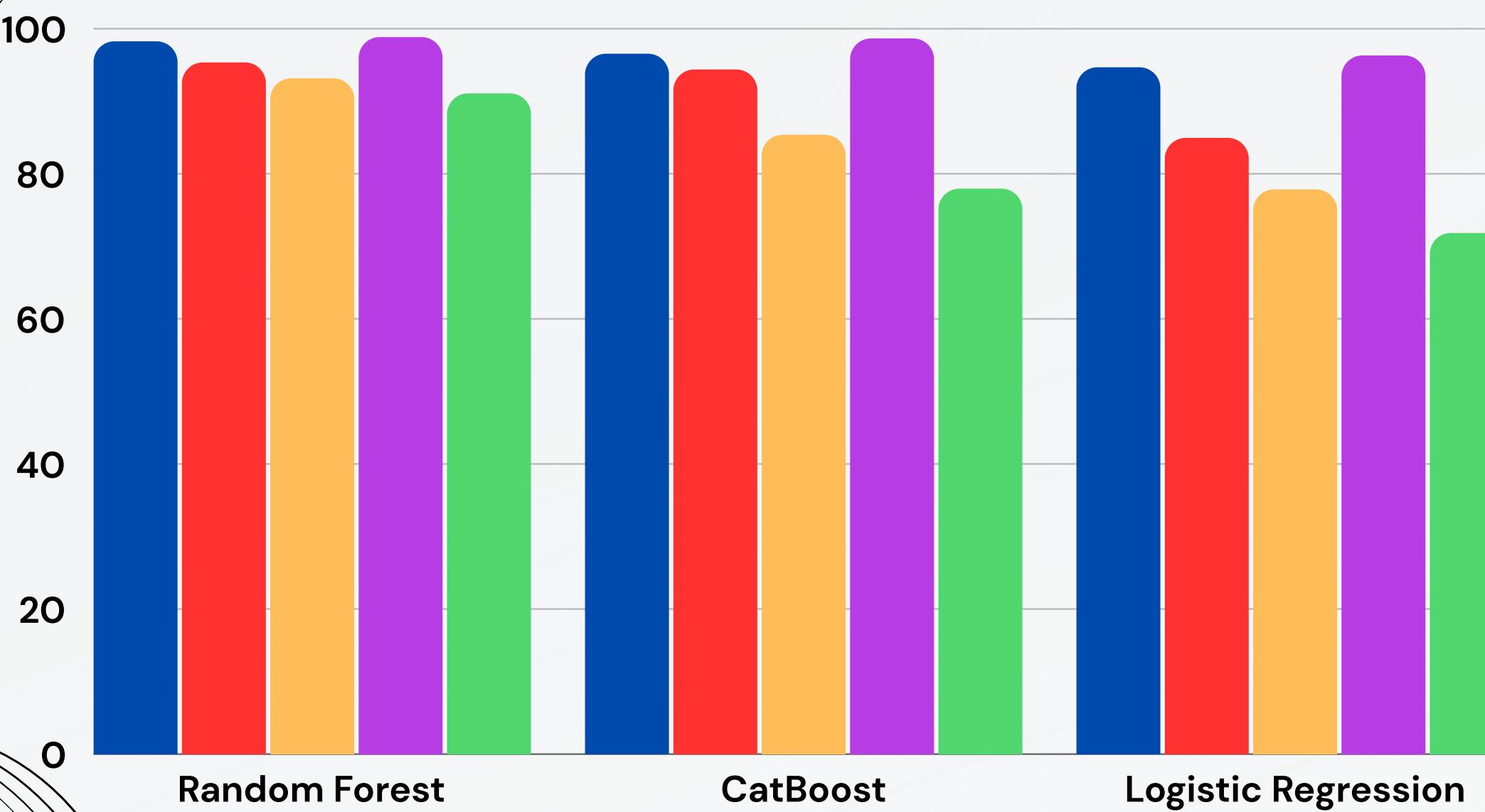
## RECALL

La proportion de cas positifs correctement identifiés sur le total des cas positifs réels, mesurant la capacité du modèle à identifier les cas positifs.

# RÉSULTATS



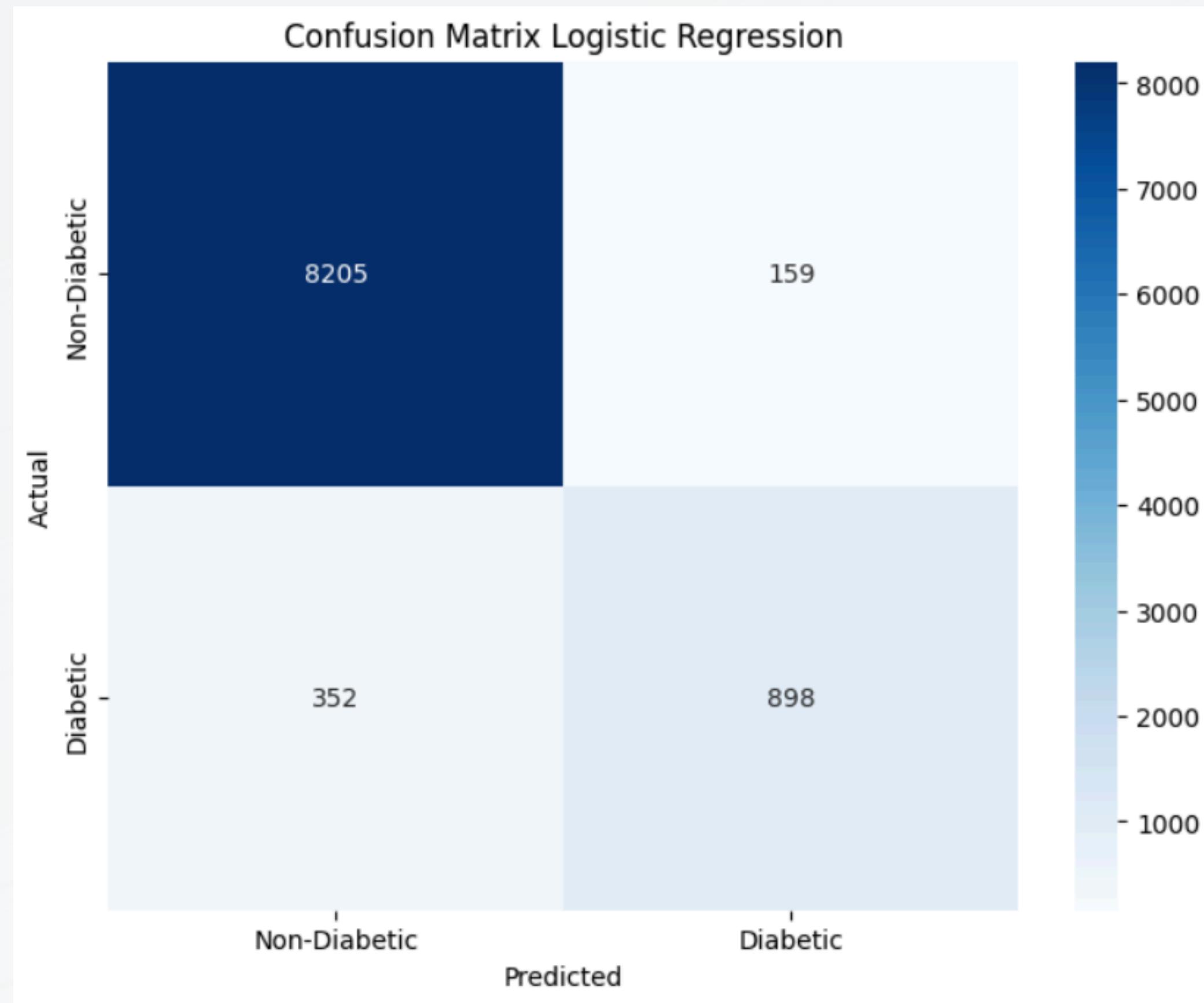
# PERFORMANCES



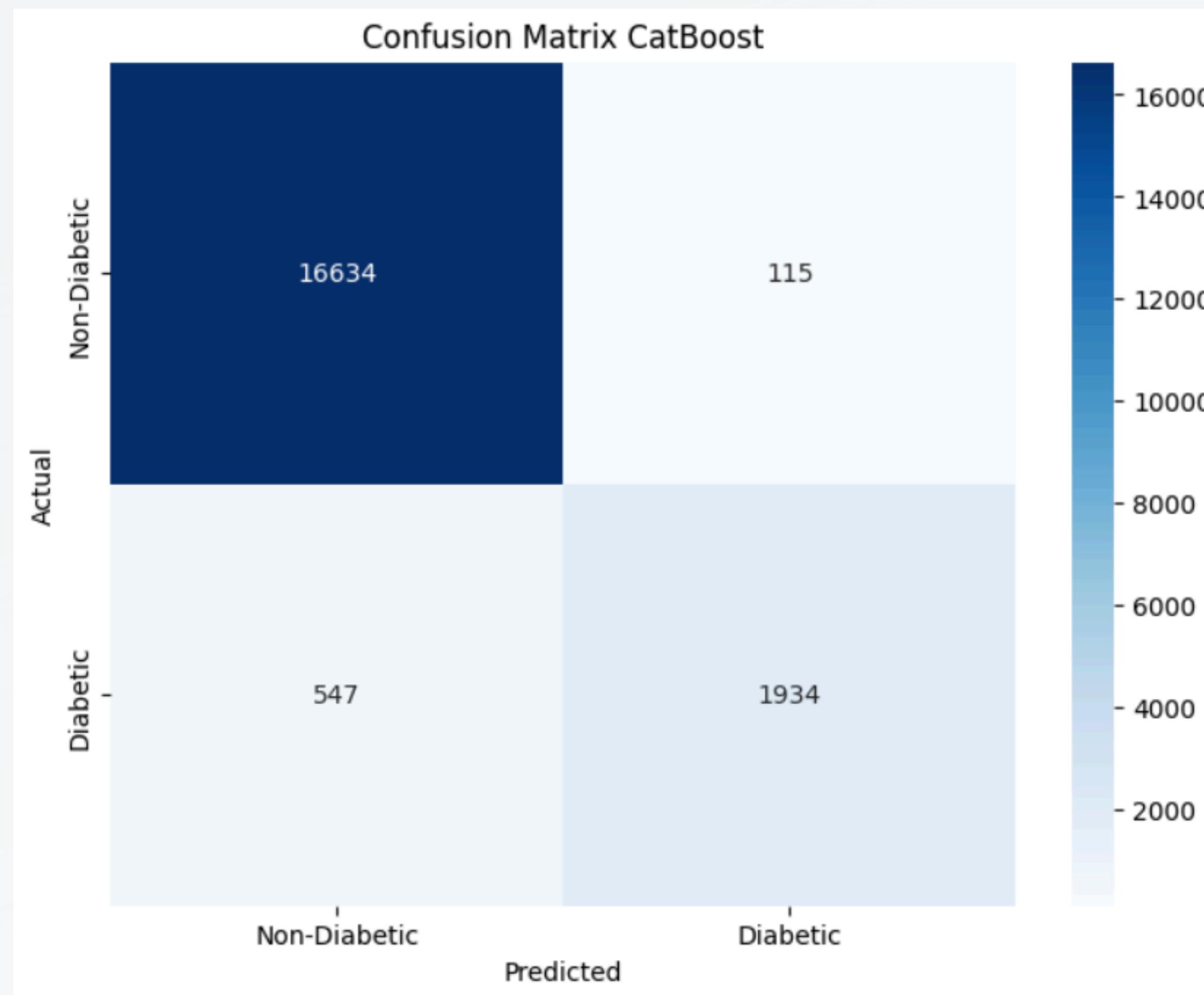
- Random Forest :
  - Accuracy : 98.26%
  - Precision : 95.35%
  - F1-Score : 93.16%
  - ROC AUC : 98.84%
  - Recall : 91.08%
- CatBoost :
  - Accuracy : 96.557%
  - Precision : 94.387%
  - F1-Score : 85.38%
  - ROC AUC : 98.66%
  - Recall : 77.95%
- Logistic Regression :
  - Accuracy : 94.68%
  - Precision : 84.95%
  - F1-Score : 77.85%
  - ROC AUC : 96.31%
  - Recall : 71.84%



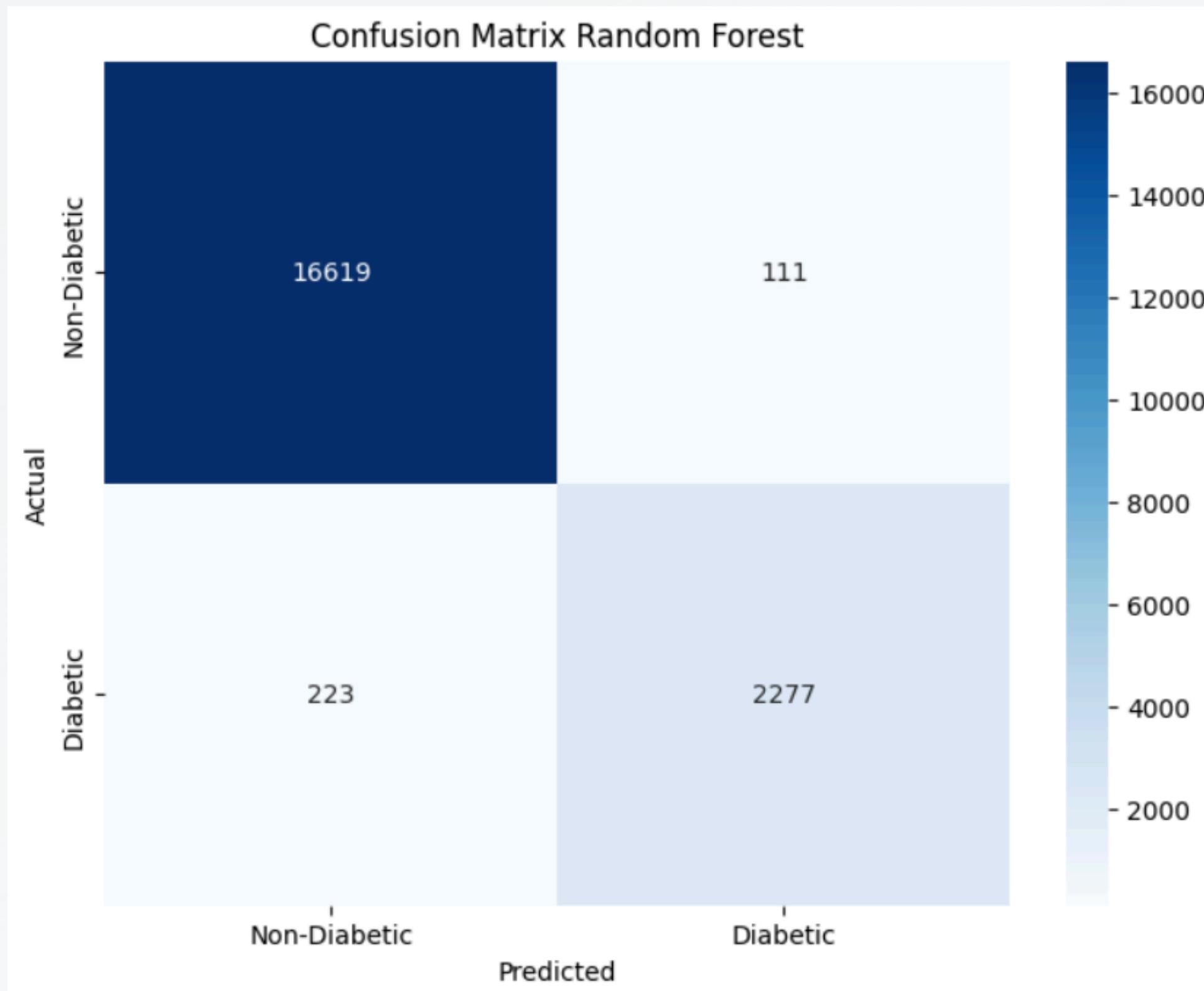
# MATRICE DE CONFUSION



# MATRICE DE CONFUSION



# MATRICE DE CONFUSION

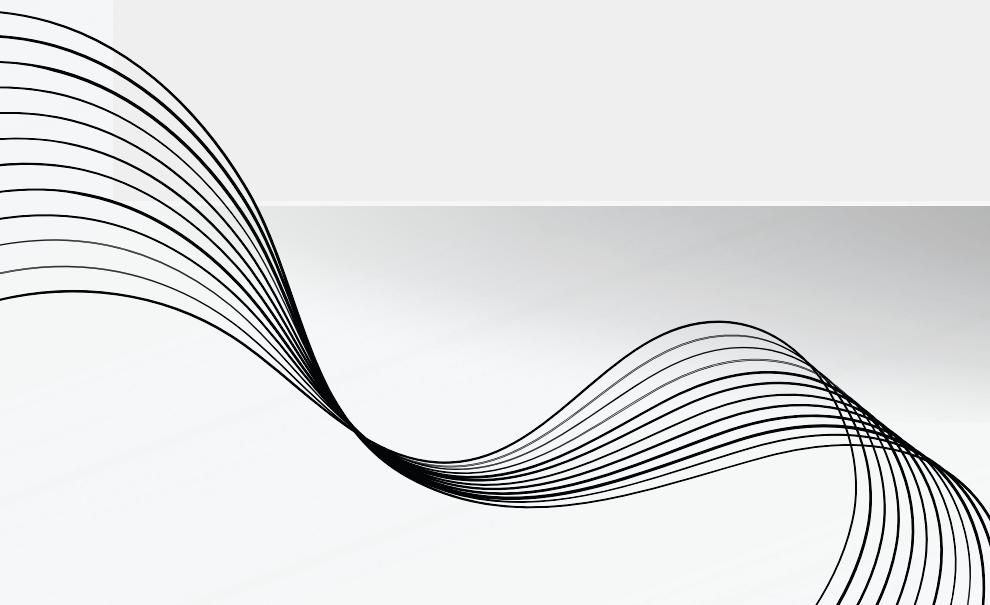


Matrice de confusion Random Forest

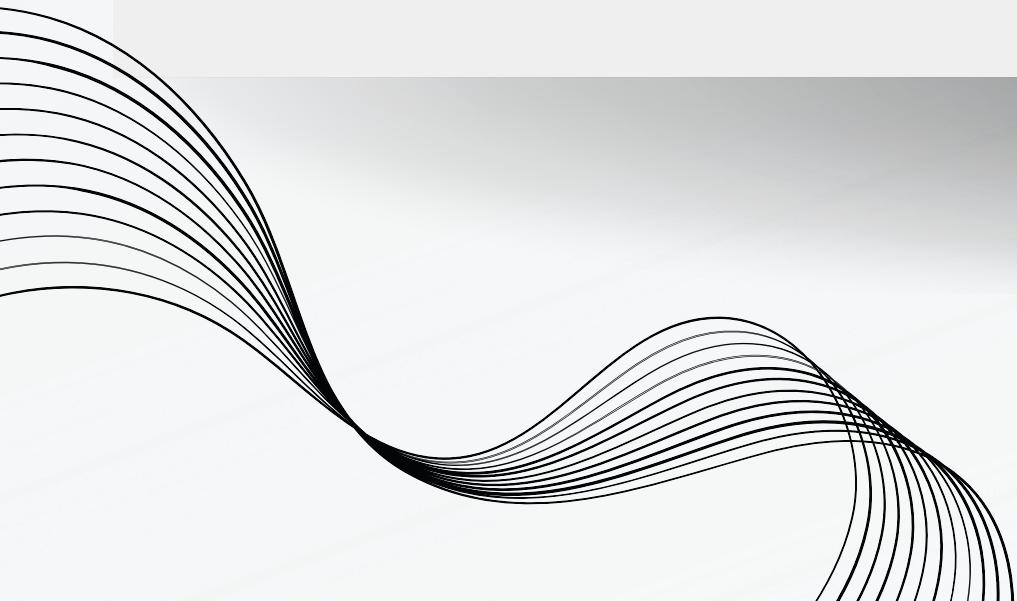
# INTERPRÉTATION DES RÉSULTATS

## RÉSULTATS APRÈS L'ENTRAÎNEMENT

- Random Forest a montré les meilleures performances parmi les modèles testés.
- Sa capacité à capturer des relations complexes et à gérer des jeux de données déséquilibrés explique ces résultats.



# CONCLUSION



# RÉFÉRENCES

- <https://www.coursera.org/learn/ibm-exploratory-data-analysis-for-machine-learning/>
- <https://youtube.com/@Deeplearningai> (Andrew Ng)

**MERCI POUR  
VOTRE ATTENTION**

