جامعة سيدي محمد بن عبد الله بفاس
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ ⴱⴼⴰⵙ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

كلية العلوم والتقنيات فاس
ⵜⴰⵖⵢⴻⵍⵜ ⵏ ⵜⵜⵛⵓⵓⵏⵉⵏ ⴰ ⵜⵓⵏⵖⵣⵜⵉ - ⵅⴰⵓ
FACULTÉ DES SCIENCES ET TECHNIQUES DE FÈS

Comparaison des Techniques d'Apprentissage Supervisé pour la Prédiction du Diabète

ANALYSE DES PERFORMANCES DE KNN, SVM, DT ET RF SUR LE DATASET "DIABETES PREDICTION"

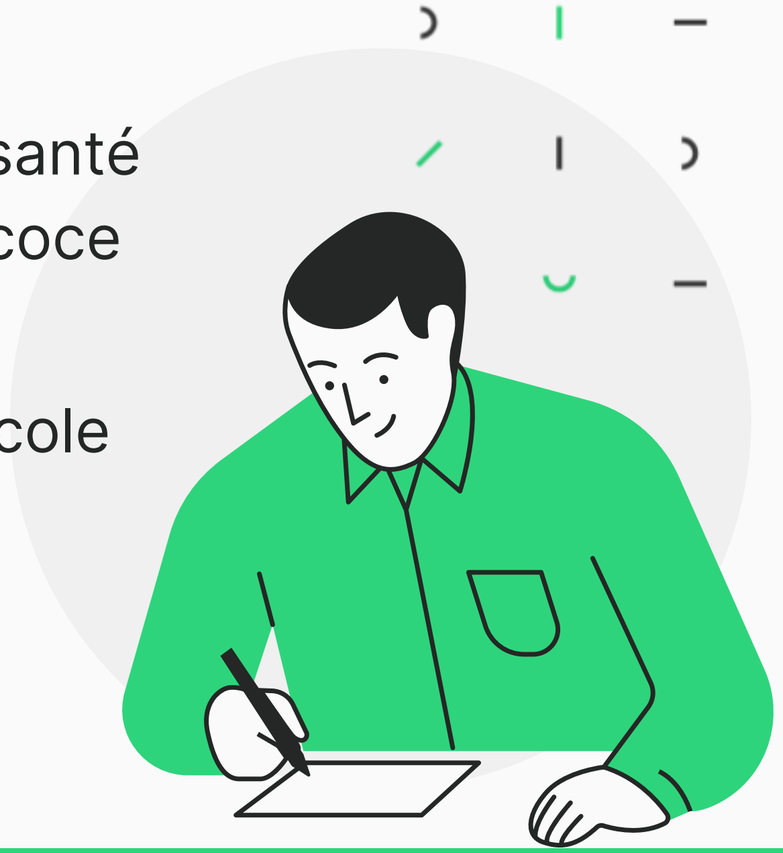
ELABORÉ PAR :
EL AMRAOUI MOHAMED
ENCADRÉ PAR :
PR. JAMAL KHARROUBI

Plan du Rapport

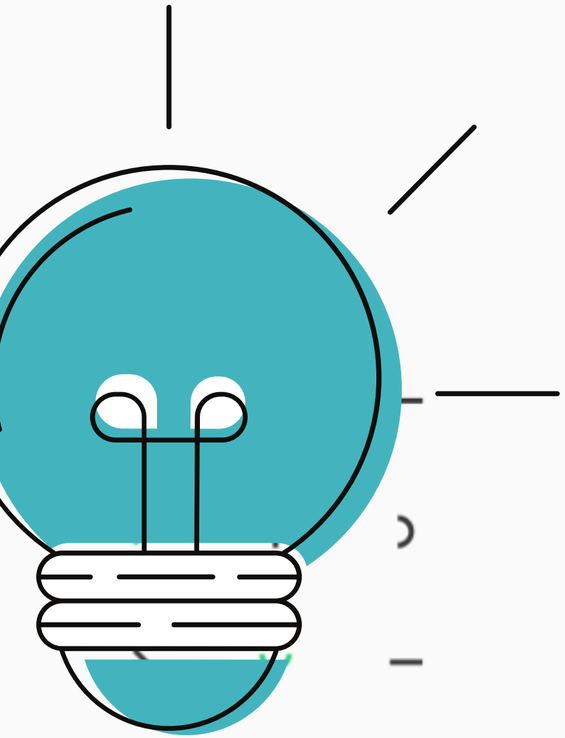
- **1- Plan du Rapport :** Aperçu des sections à venir.
- **2-Introduction :** Objectif du projet et importance du sujet.
- **3-Justification du Choix du Dataset :** Raisons du choix et lien Kaggle.
- **4-Description du Dataset :** Détails sur les colonnes, statistiques clés.
- **5-Protocole Expérimental :** Étapes suivies pour l'analyse.
- **6-Résultats - Comparaison des Performances :** Tableau comparatif (KNN, SVM, DT, RF).
- **7-Conclusion et Discussion :** Résumé, limites, améliorations possibles.
- **8-Références et Accès au Code :** Sources et lien GitHub pour le code.

Introduction

- **Objectif :** Comparer les performances de KNN, SVM, DT et RF pour prédire le diabète à partir de données médicales.
- **Importance :** Le diabète, un enjeu majeur de santé publique, peut bénéficier d’une détection précoce via des modèles prédictifs.
- **Aperçu :** Choix du dataset, description, protocole expérimental, résultats, conclusion.
- **Code :** Disponible sur GitHub (lien en fin de présentation).

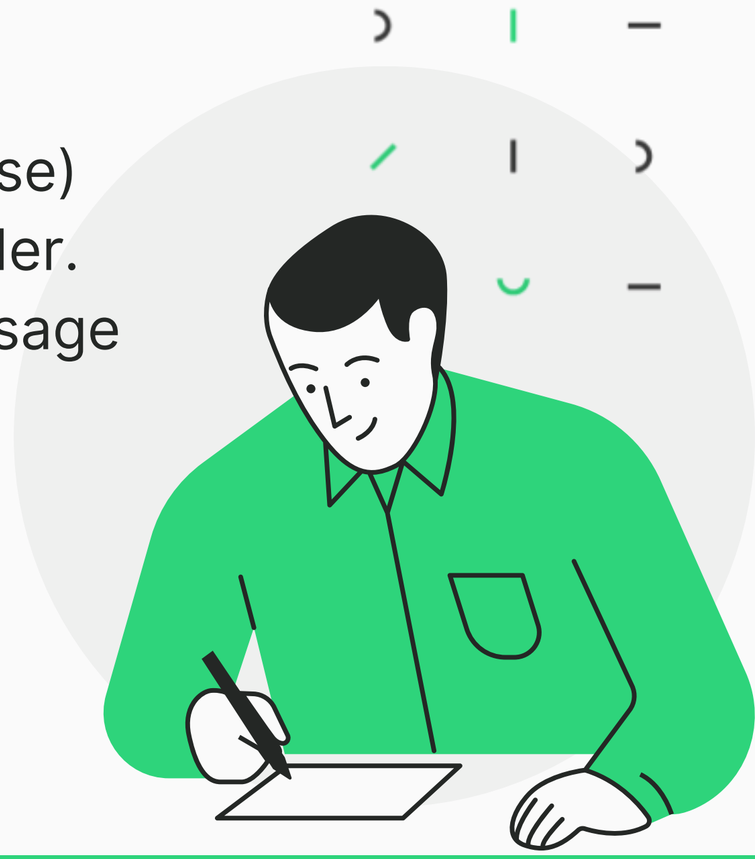


Dataset : Choix et Description



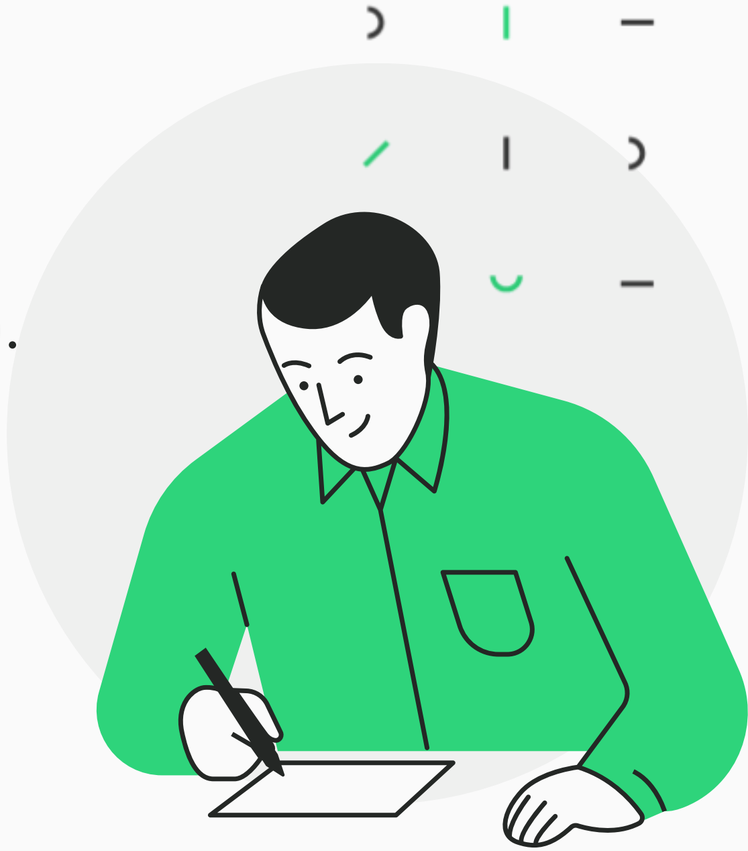
Justification du Choix du Dataset

- **Lien :** Diabetes Prediction Dataset - Kaggle
- **Pourquoi ce dataset ?**
  - **Classification binaire :** Prédire le diabète (0 ou 1), idéal pour KNN, SVM, DT, RF.
  - **Structure :** Données numériques (âge, IMC, glucose) + catégoriques (gender, smoking_history) à encoder.
  - **Taille :** 100 000 lignes, adaptée pour un apprentissage efficace sans temps excessif.
  - **Pertinence** : Problème médical réel, utile pour comparer les algorithmes (accuracy, F1-score).
- **Conclusion :** Dataset simple, riche, et pertinent pour une analyse claire et impactante.



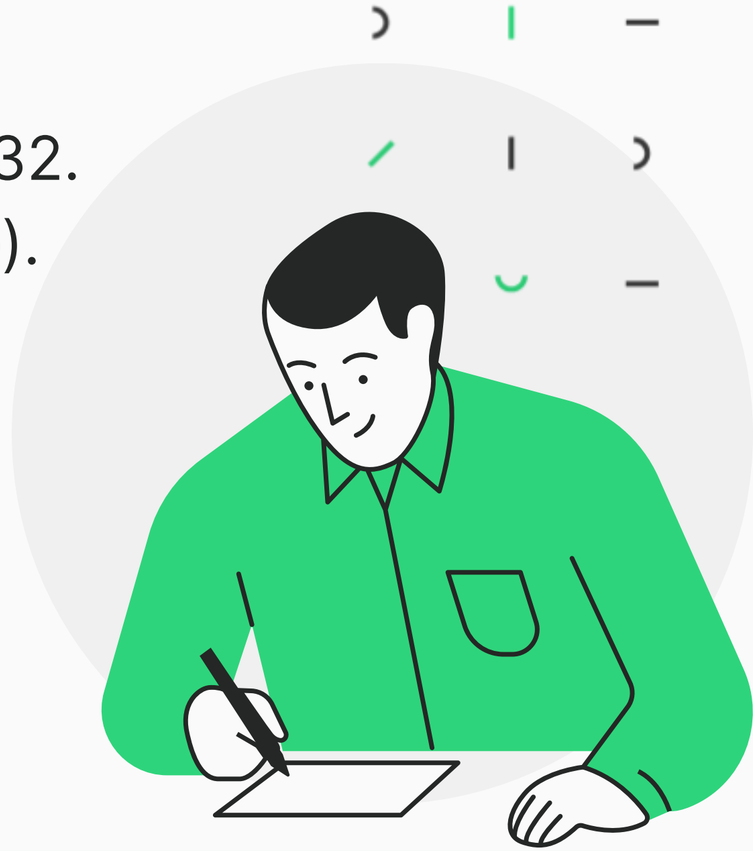
Description du Dataset

- **Dataset :** "Diabetes Prediction" (données médicales et démographiques).
- **Colonnes (9) :**
  - **Âge :** 0-80 ans (risque accru avec l’âge).
  - **Genre :** Homme, Femme, Autre.
  - **IMC :** 10.16-71.55 (lié au diabète).
  - **Hypertension, Maladie cardiaque :** 0 (non), 1 (oui).
  - **Historique de tabagisme :** Non-fumeur, ancien, actuel, etc.
  - **HbA1c :** 3.5-9% (sucre sanguin).
  - **Glucose :** 80-300 mg/dL (signe de diabète).
  - **Diabète :** Cible (0 = non, 1 = oui).



Description du Dataset

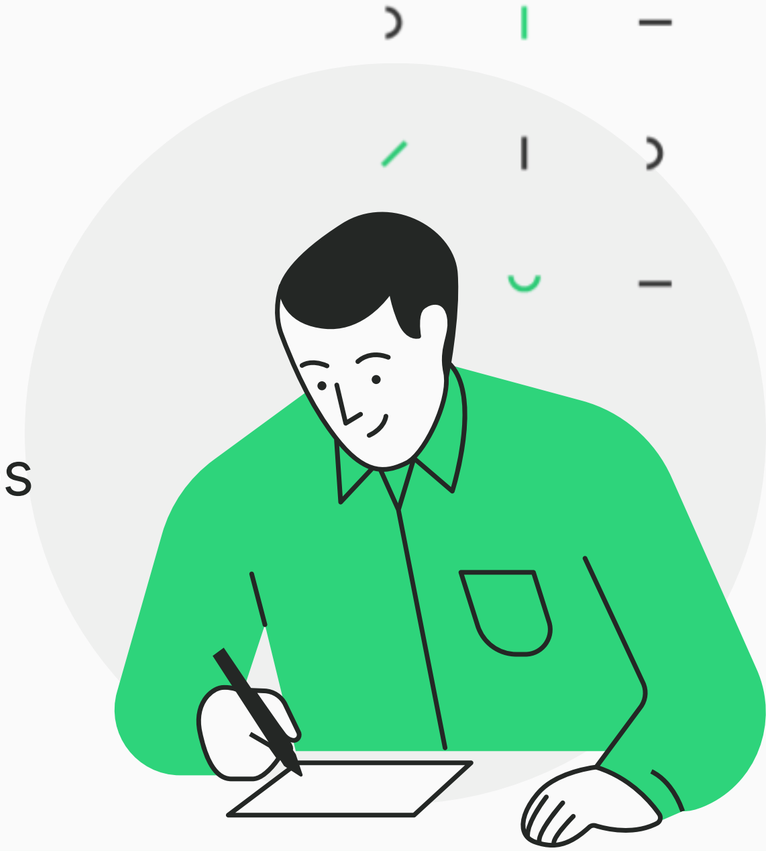
- **Stats clés :**
  - **Taille :** 3.81 MB, 100 000 lignes.
  - **Âge moyen :** 41.89 ans, IMC moyen : 27.32.
  - **Prévalence diabète :** 8.5 % (déséquilibré).
- **Utilité :** Classification binaire pour la santé.



Protocole Expérimental

Protocole Expérimental (1/2)

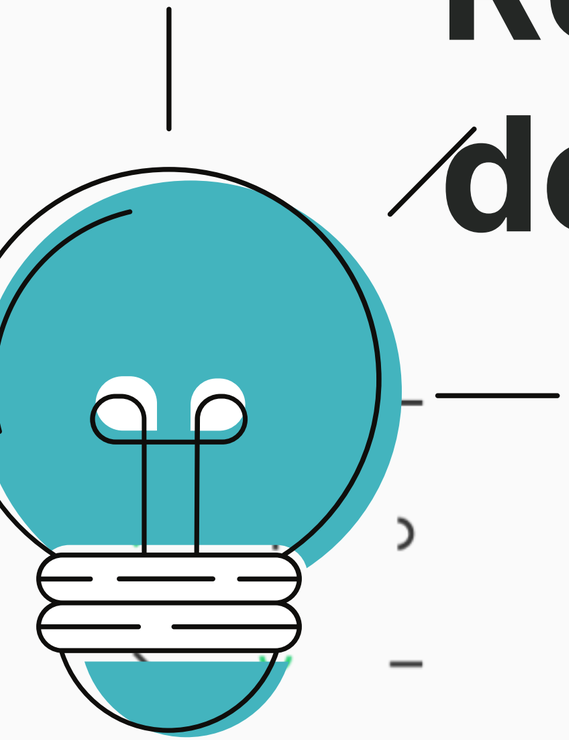
- **Étapes pour analyser le dataset "Diabetes Prediction" :**
  - **Description :** Analyse de la structure et des valeurs manquantes (aucune détectée).
  - **Encodage :** One-Hot Encoding pour gender et smoking_history.
  - **Outliers :** Détection (boxplots/Z-scores) et suppression (ex. : IMC, glucose).
  - **Corrélation :** Heatmap pour identifier les variables influentes (ex. : glucose, HbA1c).
  - **Distribution cible :** Déséquilibre (91.5 % de 0, 8.5 % de 1).



Protocole Expérimental (2/2)

- **Suite des étapes :**
    - **6. Scaling :** StandardScaler pour uniformiser les échelles (essentiel pour KNN/SVM).
    - **7. Cross-validation** : K-Fold (K=5) pour des performances robustes.
    - **8. Application des modèles :** KNN, SVM, DT, RF + visualisation (matrice de confusion, PCA) + évaluation (accuracy, F1, rappel).
- **Échantillonnage (10 %) :**
    - **Raison :** Réduit le temps élevé (~100 000 lignes) .
    - **Impact :** Analyse rapide, mais limite la généralisation.

Résultats - Comparaison des Performances



- **Comparaison des Performances (Échantillon 10 %)**

Modèle	Accuracy (Cross-Val)	F1-Score (Cross-Val)	Précision (Cross-Val)	Rappel (Cross-Val)	Accuracy (Val)	(Val)F1-Score (Val)	Précision (Val)	Rappel (Val)	Temps de Calcul (s)
KNN	0.958	0.700 (±0.037)	0.899 (±0.013)	0.575 (±0.051)	0.97	0.78	0.96	0.66	11.14
SVM	0.963	0.731 (±0.014)	0.971 (±0.009)	0.587 (±0.017)	0.97	0.75	0.98	0.61	126.65
DT	0.952 (±0.004)	0.729 (±0.017)	0.714 (±0.024)	0.747 (±0.043)	1.00	1.00	1.00	1.00	0.81
RF	0.972 (±0.002)	0.813 (±0.014)	0.984 (±0.009)	0.693 (±0.017)	1.00	1.00	1.00	1.00	16.27

**** Métriques : Les métriques de validation (Accuracy, F1-Score, Précision, Rappel) sont extraites des rapports de classification pour la classe 1 (diabète positif).***

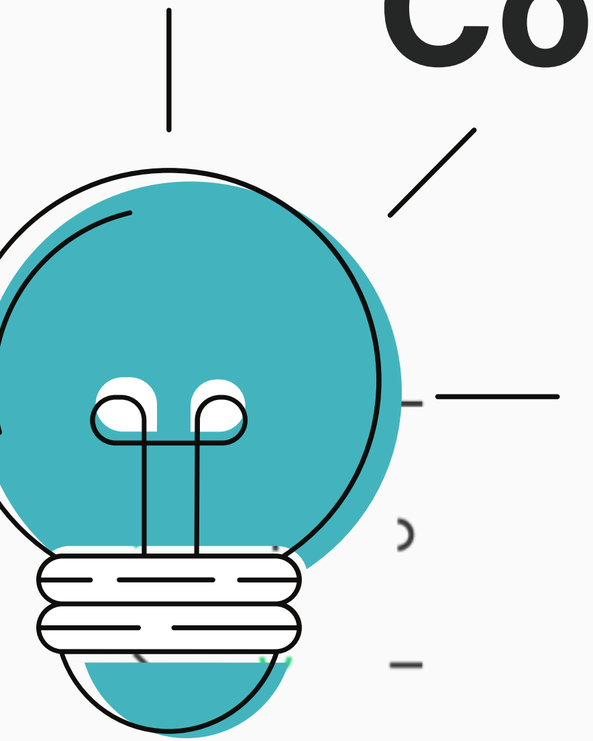
Résultats - Comparaison des Performances

- **Observations :**
  - **DT et RF dominent en validation :** DT et RF atteignent des métriques parfaites (Accuracy, F1, Précision, Rappel 1.00) en validation, mais cela indique un surapprentissage, car leurs performances Cross-Val sont inférieures (ex. : F1 0.729 pour DT, 0.813 pour RF).
  - **RF excelle en Cross-Val :** RF surpasse les autres modèles en Cross-Val (Accuracy 0.972, F1 0.813, Précision 0.984), mais son rappel (0.693) est inférieur à DT (0.747), montrant une détection moins complète de la classe 1.

Résultats - Comparaison des Performances

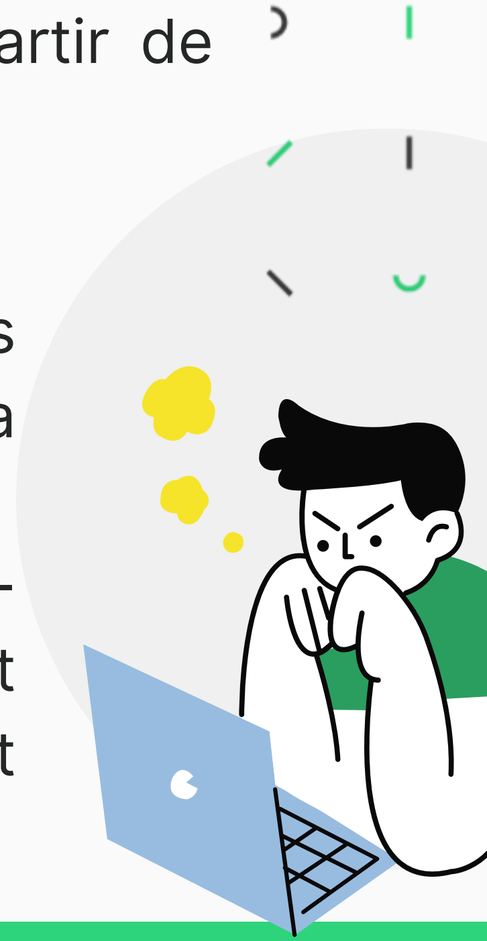
- **Observations :**
  - **KNN et SVM performants en validation :** KNN (F1 0.78, Rappel 0.66) et SVM (F1 0.75, Rappel 0.61) affichent de bonnes performances en validation, mais leur rappel reste limité, indiquant une difficulté à détecter la classe minoritaire (1).
  - **Temps de calcul :** DT est le plus rapide (0.81 s), suivi de KNN (11.14 s) et RF (16.27 s). SVM est très lent (126.65 s), ce qui le rend moins pratique pour des datasets plus grands.
  - **Déséquilibre des classes :** Le déséquilibre (91.5 % 0, 8.5 % 1) impacte les performances, surtout en Cross-Val, où KNN et SVM ont un rappel plus faible (0.575 et 0.587).

Conclusion et Discussion



Conclusion et Discussion

- **Résumé**
  - Ce projet a permis de comparer quatre algorithmes de machine learning (KNN, SVM, DT, RF) pour prédire le diabète à partir de données médicales et démographiques.
  - **Performances :**
    - En validation : DT et RF atteignent des scores parfaits (accuracy, F1-score, précision, rappel = 1.00), mais cela indique un surapprentissage.
    - En validation croisée : RF domine (accuracy 0.972, F1-score 0.813), suivi de SVM (F1-score 0.731), DT (0.729) et KNN (0.700). Le rappel reste limité pour KNN (0.575) et SVM (0.587).



Conclusion et Discussion

- **Résumé**

  - Ce projet a permis de comparer quatre algorithmes de machine learning (KNN, SVM, DT, RF) pour prédire le diabète à partir de données médicales et démographiques.
  - **Performances :**
    - En validation : DT et RF atteignent des scores parfaits (accuracy, F1-score, précision, rappel = 1.00), mais cela indique un surapprentissage.
    - En validation croisée : RF domine (accuracy 0.972, F1-score 0.813), suivi de SVM (F1-score 0.731), DT (0.729) et KNN (0.700). Le rappel reste limité pour KNN (0.575) et SVM (0.587).

Conclusion et Discussion

- **Résumé**
  - **Temps de calcul:**
    - DT est le plus rapide (0.81 s), suivi de KNN (11.14 s) et RF (16.27 s). SVM est très lent (126.65 s), peu pratique pour des datasets volumineux.
    - L’échantillonnage à 10% du dataset a réduit le temps de calcul, permettant une analyse faisable.

Conclusion et Discussion

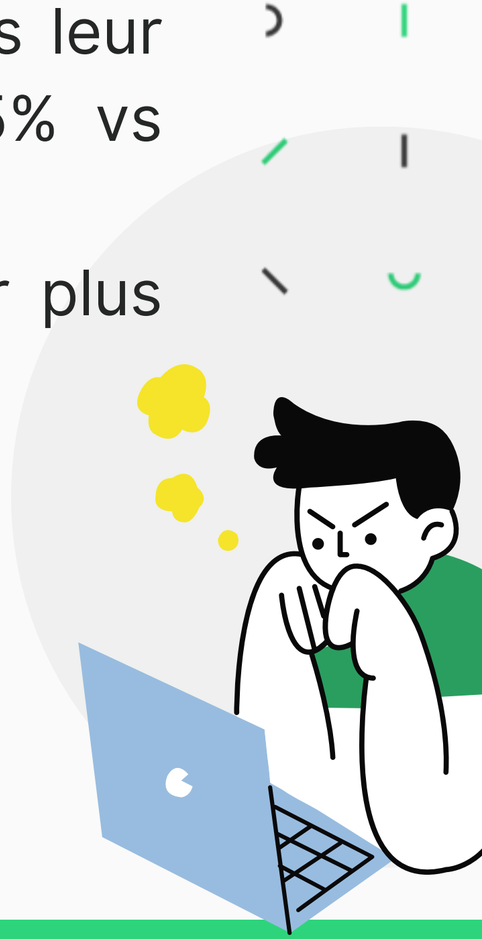
- **Limites**
  - **Surapprentissage :** DT et RF excellent en validation (scores de 1.00), mais leurs performances en validation croisée sont inférieures, signe d'une faible généralisation.
  - **Déséquilibre des classes :** Avec 91.5% de non-diabétiques et 8.5% de diabétiques, les modèles (surtout KNN et SVM) peinent à détecter la classe minoritaire (rappel faible).
  - **Échantillonnage réduit :** L'utilisation de 10% du dataset (10 000 lignes) peut introduire un biais et limiter la robustesse des résultats.
  - **Séparation des classes :** L'ACP montre une variance expliquée de 48.2%, indiquant une séparation limitée entre diabétiques et non-diabétiques.

Conclusion et Discussion

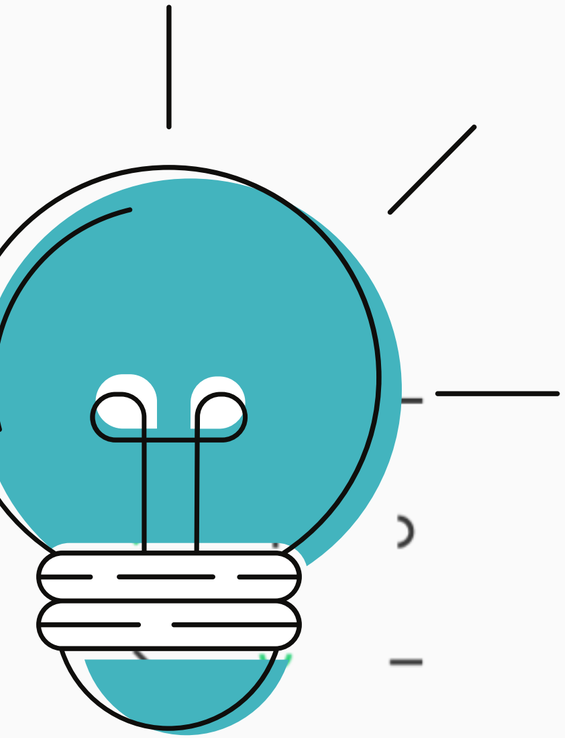
- **Améliorations possibles**
  - **Optimisation :** Utiliser GridSearchCV/RandomizedSearchCV (ex. : max_depth pour DT, n_estimators pour RF) pour réduire le surapprentissage.
  - **Déséquilibre :** Appliquer SMOTE pour la classe minoritaire ou utiliser l'AUC-ROC comme métrique.
  - **Dataset complet :** Entraîner sur 100 000 lignes pour plus de robustesse, si possible.
  - **Feature engineering :** Créer des variables (ex. : âge-glucose) et améliorer la réduction de dimension.
  - **Nouveaux algorithmes :** Tester Gradient Boosting ou XGBoost pour surpasser RF.
  - **Validation croisée :** Adopter une approche stratifiée pour équilibrer les classes.

Conclusion finale

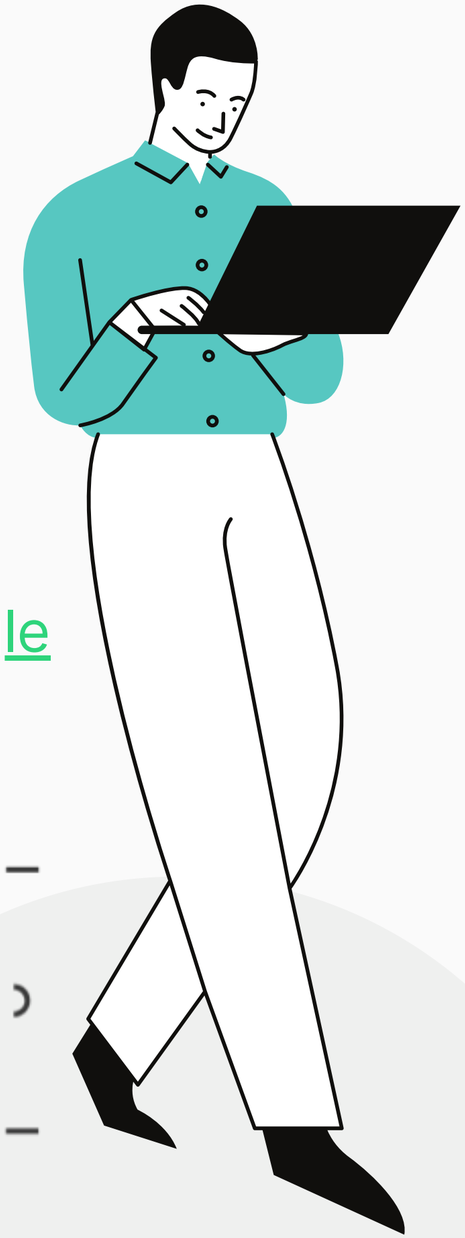
- DT et RF excellent en validation (scores parfaits), mais leur surapprentissage et le déséquilibre des classes (91.5% vs 8.5%) limitent leur généralisation.
- Avec ces ajustements, les modèles pourraient devenir plus fiables pour prédire le diabète en contexte médical.



Références et Accès au Code



Références et Accès au Code



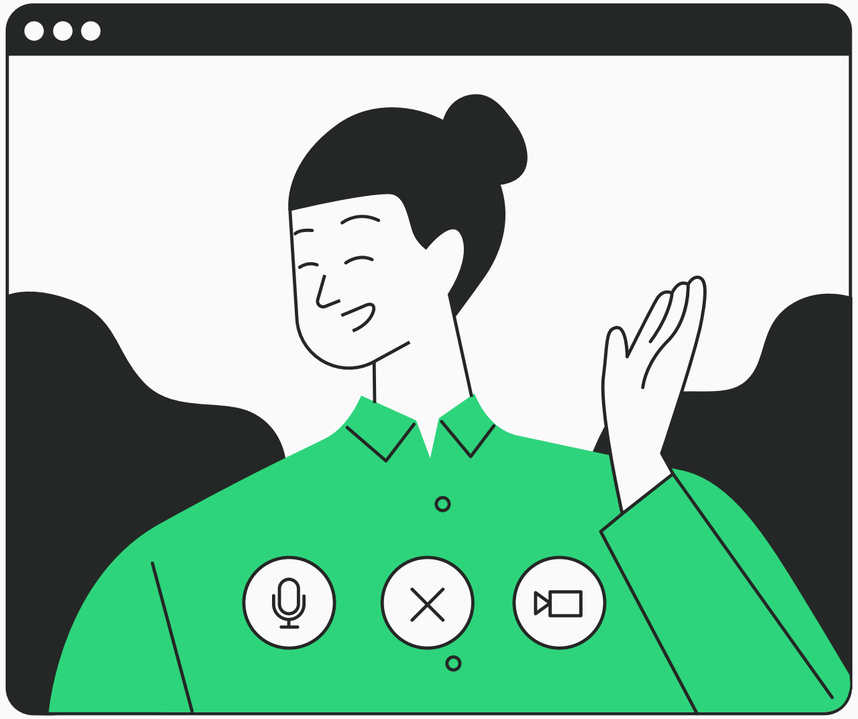
KAGGLE

- Lien : Diabetes Prediction Dataset - Kaggle

GITHUB

- Lien :Predicting-Diabetes-with-ML

Contact me



EMAIL

mohamedelamrawi@yahoo.com

WEBSITE

https://linktr.ee/el_amraoui_mohamed

PHONE

+212712542184