Université TÉLUQ

Certificat en sciences des données

TITRE DU PROJET :

PRÉDICTION DU RISQUE DE DIABÈTE À L'AIDE DE L'APPRENTISSAGE AUTOMATIQUE : ÉTUDE EXPLORATOIRE AVEC LE JEU DE DONNÉES « DIABETES PREDICTION DATASET »

Cours : SCI1402 - Projet en sciences de données

Projet fonctionnel : Projet Final

Nom de l'étudiant : Orly Steve Ngayap Tchouamdjou

Numéro d'étudiant : 25134271

Adresse courriel : ngayap_tchouamdjou.orly_steve@univ.teluq.ca

Tuteur : Fatima Bensalma

Session : Automne 2025

Date de remise : 09 décembre 2025

1. Rappel du projet et des objectifs

Le projet vise à développer un modèle de classification prédictive capable d'estimer le risque de diabète à partir de caractéristiques biométriques et comportementales issues du jeu de données *« Diabetes Prediction Dataset »* du site www.kaggle.com

Selon la ***World Health Organization (WHO, 2023);*** la prédiction précoce du diabète permet de réduire les complications cardio-vasculaires et la mortalité.

L'objectif général est de comparer plusieurs algorithmes d'apprentissage automatique (régression logistique, KNN, forêts aléatoires, SVM) et d'évaluer leur capacité à distinguer les individus diabétiques des non diabétiques à l'aide des métriques comme l'AUC-ROC, la précision, le rappel et le F1-score.

Le présent rapport mi-parcours décrit l'état d'avancement du projet, les analyses déjà réalisées sur les données ainsi que la stratégie de modélisation prévue pour la suite.

2. Description du jeu de données et préparation

2.1. Importation du jeu de données et structure

Ce que nous avons fait : Importer le fichier CSV, vérifier le nombre de lignes, de colonnes et le type des variables (Voir fichier RMarkdown).

Commentaire : Le jeu de données comporte **100 000 observations** et **9 variables**. Les variables disponibles sont : `gender`, `age`, `hypertension`, `heart_disease`, `smoking_history`, `bmi`, `HbA1c_level`, `blood_glucose_level` et la variable cible `diabetes`. La commande `str()` confirme que certaines variables sont numériques (`age`, `bmi`, `HbA1c_level`, `blood_glucose_level`), alors que d'autres sont codées sous forme de texte ou d'entiers binaires (0/1).

2.2. Description détaillée des variables

Après inspection dans R, les variables du jeu de données sont décrites comme suit :

- **`gender`** : variable catégorielle indiquant le sexe de la personne (Female, Male, Other)
- **`age`** : variable numérique continue représentant l'âge en années.
- **`hypertension`** : variable binaire (0/1) indiquant la présence d'hypertension (1 = oui, 0 = non).
- **`heart_disease`** : variable binaire (0/1) indiquant la présence d'une maladie cardiaque
- **`smoking_history`** : variable catégorielle décrivant les habitudes tabagiques (never, current, former, ever, not current, No Info).
- **`bmi`** : variable numérique continue représentant l'indice de masse corporelle
- **`HbA1c_level`** : variable numérique continue correspondant au taux d'hémoglobine glyquée, un biomarqueur important du diabète. L'HbA1c est un indicateur reconnu du contrôle glycémique sur plusieurs semaines, selon (ADA, 2023).

- **`blood_glucose_level`** : variable numérique continue représentant le niveau de glucose sanguin.
- **`diabetes`** : variable binaire (0/1) faisant office de variable cible : 1 = individu diabétique, 0 = individu non diabétique.

Cette description permet de clarifier le rôle et la nature de chaque variable avant la modélisation.

2.3. Vérification des valeurs manquantes et des doublons

➢ **Vérification**

A ce niveau nous allons vérifier s'il existe des valeurs manquantes (NA) et des lignes dupliquées (Voir code RMarkdown).

➢ **Interprétation** : L'analyse de la qualité des données nous montre que :

- **Aucune valeur manquante** n'est présente dans le jeu de données `(na_total =0)`. Cela signifie que toutes les variables sont complètes pour les 100 000 observations, ce qui facilite les étapes de préparation et évite la nécessité d'utiliser des méthodes d'imputation ou de suppression d'observations.
- En revanche, l'analyse des doublons indique la présence de **3854 lignes** strictement identiques au sein du dataset `(nb_doublons = 3854)`. Ce volume représente environ **3.85%** de l'ensemble des données. La présence de doublons peut introduire un biais dans les analyses descriptives et dans l'apprentissage des modèles, car certains profils seraient artificiellement surreprésentés.

Ainsi, même si le dataset est complet et sans valeurs manquantes, une étape de suppression des doublons sera nécessaire pour garantir la qualité statistique des analyses et la fiabilité des futures modèles prédictifs.

➢ **Suppression des doublons**

Pour le faire nous allons créer un nouveau jeu de données sans les lignes dupliquées, puis vérifier la nouvelle dimension du dataset. (Voir RMarkdown).

Commentaire : Après suppression des doublons, le nombre d'observations est passé de **100 000 à 96 146**, ce qui confirme bel et bien la suppression de **3 854 lignes** identifiées précédemment. Le jeu de données nettoyé `(diabetes_clean)` ne contient désormais plus aucun doublon, ce qui garantit que chaque ligne représente un individu unique. Cette étape renforce la fiabilité des analyses statistiques et des futurs modèles prédictifs, en évitant la surreprésentation artificielle de certains profils.

2.4. Conversion des types de variables

Après les différentes étapes de vérification de la clarté des données, ce qui a permis de supprimer les doublons, nous devons convertir les variables catégorielles et binaires en facteurs. (Voir RMarkdown).

Commentaire : L'analyse de la structure du jeu de données après conversion confirme que les variables ont désormais les types appropriés. Cette conversion est essentielle pour garantir le bon fonctionnement des analyses statistiques (comme les tableaux croisés) et des modèles

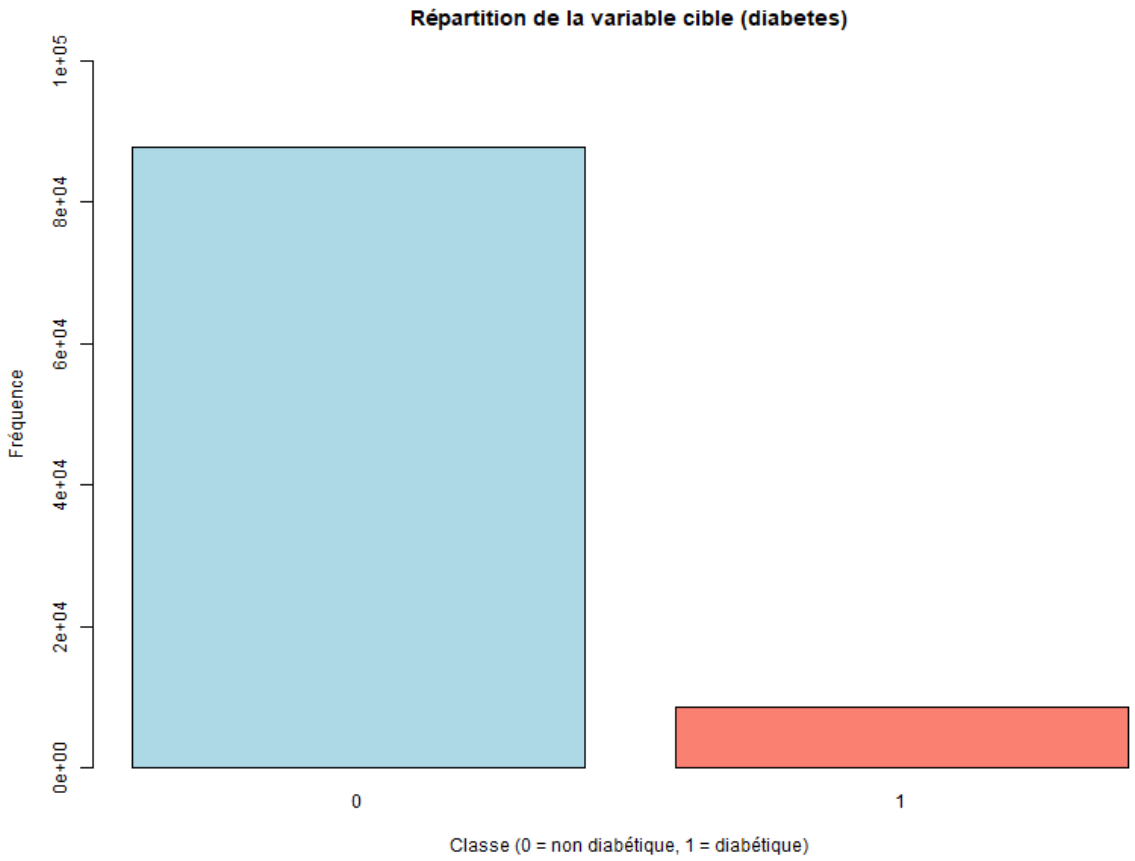
d'apprentissage supervisé, qui exploitent différemment les variables numériques et catégorielles.

Le jeu de données est maintenant propre, complet, sans doublons et correctement typé, ce qui permet d'aborder sereinement l'analyse exploratoire.

3. Analyse exploratoire

3.1. Distribution de la variable cible `diabetes`

A ce niveau nous allons calculer les fréquences des classes (0 et 1), calculer les proportions et produire un graphique en barres pour visualiser le déséquilibre de classes. (Voir RMarkdown)



Interprétation : La variable cible `diabetes` présente un déséquilibre marqué entre les deux classes. Après nettoyage des doublons, le jeu de données contient : une très forte majorité d'individus non diabétiques (classes 0), et une proportion beaucoup plus faible d'individus (classe 1).

Le graphique en barres permet de visualiser ce déséquilibre; la classe 0 regroupe la quasi-totalité des observations, tandis que la classe 1 représente une minorité d'individus. Ce déséquilibre des classes est typique des problèmes de prédiction médicale. Il est important, car il peut affecter les performances des modèles : un modèle naïf pourrait obtenir une bonne précision globale en prédisant systématiquement la classe 0.

Pour cette raison, il sera nécessaire d'utiliser des métriques adaptées (Recall, F1-score, AUC-ROC) ou des stratégies spécifiques telles que le sur-échantillonnage lors de la phase de modélisation.

3.2. Statistiques descriptives des variables numériques

Nous allons maintenant résumer les variables numériques, identifier les moyennes, médianes et valeurs extrêmes et enfin préparer l'analyse pour les comparaisons entre classes. (Voir RMarkdown).

Commentaire/Interprétation : Les variables numériques du jeu de données présentent les caractéristiques suivantes :

- ✓ **Âge (age)**
  - Min = 0.08; Max = 80
  - Moyenne = 41.79; Médiane = 43
  - Les quartiles (24 et 59 ans) montrent une répartition large et équilibrée. Cela confirme la présence d'individus jeunes, adultes et âgés, ce qui est favorable pour la modélisation.
- ✓ **Indice de masse corporelle (bmi)**
  - Moyenne = 27.32 ; Médiane = 27.32
  - Premier quartile = 23.4 ; Troisième quartile = 29.86
  - Valeur maximale élevée (~ 95.7),

  Tout ceci indique des cas d'obésité sévère. Ce qui signifie que l'IMC présente une variabilité importante, cohérente avec une population générale.

- ✓ **Taux d'HbA1c (HbA1c_level)**
  - Min = 3.5 ; Max = 9
  - Moyenne = 5.53 ; Médiane = 5.8.

  La médiane et la moyenne légèrement au-dessus de la normale montrent la présence de nombreux sujets prédiabétiques ou diabétiques.

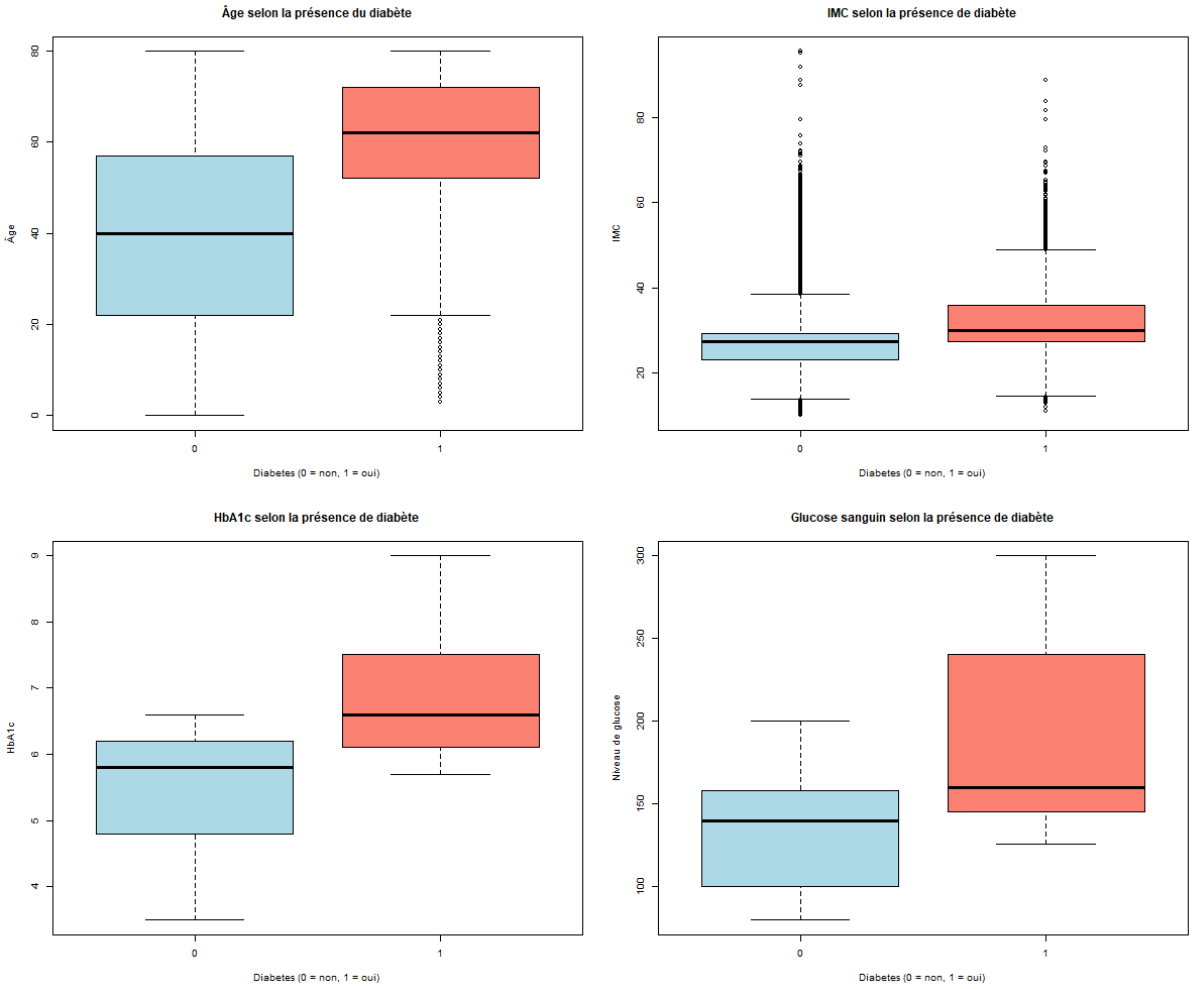
- ✓ **Glucose sanguin (blood_glucose_level)**
  - Min = 80 ; Max = 300
  - Moyenne = 138.2 ; Médiane = 140
  - Le troisième quartile = 159.

  Les niveaux de glucose montrent de fortes variations, confirmant une coexistence de profils sains et à risque.

Dans l'ensemble, les statistiques descriptives mettent en évidence une forte hétérogénéité des valeurs biologiques. Ces variations sont particulièrement importantes pour la modélisation, puisque les quatre variables numériques présentent des différences marquées entre les individus et sont connues pour contribuer fortement à la prédiction du diabète. Pour une meilleure appréciation de l'impact de chaque variable, nous allons procéder aux comparaisons des moyennes par classe.

3.3. Comparaison des moyennes par classe

Nous allons procéder au calcul des moyennes des variables numériques selon la classe diabetes et visualiser ces différences via des boxplots. (Voir RMarkdown).



<u>Interprétation :</u> Les boxplots présentés ci-dessus permettent de comparer les distributions des quatre principales variables numériques entre les individus non diabétiques (classe 0) et diabétiques (classe 1). Les résultats mettent en évidence des différences marquées entre les deux groupes :

- ✓ **Âge (age)** : Les individus diabétiques sont nettement plus âgés que les non diabétiques. Le boxplot montre : une médiane beaucoup plus élevée pour la classe 1; un écart interquartile décalé vers des valeurs supérieures et des valeurs extrêmes élevées chez les diabétiques.
  En somme, l'on constate que l'âge apparaît comme un facteur déterminant du risque de diabète, ce qui correspond aux connaissances médicales.

- ✓ **Indice de masse corporelle (bmi) :** Le bmi est systématiquement plus élevé chez les personnes diabétiques; forte différence de médiane entre les deux groupes, présence de nombreux outliers dans les deux classes mais davantage dans la classe 1, distribution globalement décalée vers des valeurs plus hautes chez les diabétiques.
  Un IMC élevé est fortement associé au diabète, ce qui confirme le jeu de données.

- ✓ **HbA1c (HbA1c_level)** : La différence est très nette; la médiane est plus élevé dans la classe diabétique, la majorité des valeurs de la classe 1 se situent dans une zone correspondant à un risque accru, aucune superposition entre les distributions des deux classes.
  On conclut que l'HbA1c est le biomarqueur le plus discriminant du dataset et reflète directement l'état glycémique chronique. « Le rôle central de l'HbA1c dans la détection du diabète est largement documenté (ADA, 2023). »

- ✓ **Glucose sanguin (blood_glucose_level)** : L'écart entre classes est également très marqué; la médiane beaucoup plus élevée chez les diabétiques, la distribution étalée et les valeurs extrêmes très hautes, peu de recouvrement entre les deux boxplots.
  Le glucose sanguin est un prédicteur très fort du diabète, comme attendu.

En somme; Les quatre variables montrent des différences claires et cohérentes entre les deux groupes. Les individus diabétiques présentent :

- un âge plus avancé,
- un IMC plus élevé,
- des taux d'HbA1c plus importants,
- un niveau de glucose sanguin nettement supérieur.

Ces observations indiquent que ces variables joueront un rôle majeur dans la modélisation et justifient leur inclusion dans les modèles prédictifs. Elles montrent aussi que le dataset est cohérent avec les connaissances médicales sur le diabète.

3.4. Hypertension et maladies cardiaques selon le diabète

Il est maintenant question calculer la proportion d'individus hypertendus selon la classe diabetes, calculer la proportion d'individus atteints de maladie cardiaque selon la classe diabetes. (Voir RMarkdown)

Commentaire/Interprétation : L'analyse des proportions d'hypertension et de maladies cardiaques selon la présence de diabète met en évidence de différences significatives entre les deux groupes :

- **Hypertension ; Classe 0 (non diabétique) : 6.13% et Classe 1 (diabétique) : 24.59%.**

  La proportion d'individus hypertendus est donc environ quatre fois plus élevée chez les personnes diabétiques que chez les non diabétiques.
  Ce résultat est cohérent avec la littérature médicale, l'hypertension étant un facteur de risque fréquent chez les individus présentant un dysfonctionnement métabolique.

- **Maladie cardiaque ; Classe 0 (non diabétique) : 3.03% et Classe 1 (diabétique) : 14.94%.**

  La prévalence des maladies cardiaques est également près de cinq plus élevée chez les individus diabétiques. Cela confirme que le diabète est fortement associé à des complications cardio-vasculaires. D'ailleurs selon la ***World Health Organization*** *(WHO, 2022);* les facteurs cardio-vasculaires comme l'hypertension sont reconnus comme augmentant le risque de diabète.

En somme, ces résultats montrent que les individus diabétiques présentent beaucoup plus souvent des antécédents d'hypertension et de maladies cardiaques que les individus non diabétiques. Ces deux variables apparaissent donc comme des facteurs de risque importants, et leur contribution sera potentiellement très forte dans les modèles prédictifs. Leur comportement est en parfaite cohérence avec les connaissances cliniques sur le diabète.

4. Première stratégie de modélisation

Nous passons maintenant à l'étape de la modélisation

4.1. Constitution du jeu d'apprentissage

Nous allons procéder à une séparation des données en deux sous-ensembles : **70% pour l'apprentissage (train_set) et 30% pour le test (test_set).** Nous utiliserons une partition stratifiée sur la variable cible `diabetes` pour conserver le même déséquilibre dans les deux sous-échantillons et vérifier ensuite la distribution de la variable cible dans chaque sous-ensemble. (Voir RMarkdown).

Commentaire : La partition a été réalisée de manière stratifiée sur la variable cible `diabetes,` de façon à conserver le même déséquilibre de classes dans les deux sous-échantillons.
Les distributions observées sont les suivantes :

- **Jeu d'apprentissage (train_set) :** Classe 0 : **61 365 observations** (~ **91,18 %**) et Classe 1 : **5 938 observations** (~ **8,82 %**)
- **Jeu de test (test_set) :** Classe 0 : **26 299 observations** (~ **91,18 %**) et Classe 1 : **2 544 observations** (~ **8,82 %**)

Les proportions des classes sont identiques dans les deux sous-ensembles, ce qui confirme que la stratification a bien fonctionné.
Cette étape garantit que les modèles entraînés ne seront pas biaisés par une distribution différente entre l'apprentissage et le test. Le déséquilibre important de la classe positive est conservé, ce qui permettra d'évaluer correctement les performances des modèles sur des données reflétant la réalité du dataset.

4.2. Premier modèle de base : régression logistique

Dans cette étape, nous ajustons un premier modèle de régression logistique afin d'estimer la probabilité de diabète à partir des variables explicatives. Ce modèle servira de référence (baseline) pour comparer les performances des modèles plus avancés. (Voir RMarkdown)

Interprétation des résultats de la régression : Le modèle de régression logistique sur le jeu d'apprentissage a été utilisé pour prédire la présence de diabète dans le jeu de test. La matrice de confusion suivante résume les performances obtenues :

- **Précision globale (Accuracy = 95.99%)** : Le modèle classifie correctement près de 96% des observations. Cependant, avec un dataset très déséquilibré (classe 1 = 8.8%), une bonne accuracy n'est pas suffisante pour juger de la qualité du modèle.
- **Sensitivity (rappel de la classe positive = 64,7%)** : La sensibilité mesure la capacité à détecter les diabétiques (classe 1). Le modèle détecte 64.7% des cas réels de diabète,

cela montre qu'il manque encore environ 35% des personnes diabétiques. C'est le point faible du modèle, ce qui est typique d'une régression logistique sur un dataset très déséquilibré.

- **Specificity (spécificité = 99.02%) :** La spécificité est très élevée : le modèle identifie correctement 99% des non-diabétiques. Cela signifie que le modèle est excellent pour prédire les personnes qui ne sont pas diabétiques.
- **Positive Predictive Value (PPV = 86.45%)** : Lorsqu'il prédit la classe 1 (diabète), il a **86%** de chances d'avoir raison. Ce qui est très élevé, et probablement dû au fait qu'il ne prédit 1 que dans des cas évidents.
- **Balanced Accuracy (0.8186)** : Come le dataset est déséquilibré, la balanced accuracy est une métrique plus fiable. Le score de 0.8186 indique une performance globalement solide, mais encore améliorable.
- **Kappa = 0.7189** : Le coefficient Kappa mesure l'accord entre les prédictions et les valeurs réelles, en tenant compte du hasard. Un Kappa de 0.7189 correspond à un bon niveau d'accord, largement supérieur à ce qui serait dû au hasard.

En conclusion, la régression logistique offre une bonne performance générale, mais montre une sensibilité insuffisante pour détecter tous les cas de diabète. Le modèle réussit très bien à identifier les non-diabétiques (spécificité élevée), mais a plus de difficulté à détecter tous les diabétiques (sensibilité faible à modérée).
Ce comportement est typique dans les problèmes de classification médicale avec classes déséquilibrées.
Ce modèle servira donc de référence (baseline) pour comparer les méthodes plus avancées (Random Forest, SVM, KNN) capables de mieux gérer le déséquilibre.

Interprétation des coefficients du modèle logistique : Les résultats obtenus présentent les coefficients estimés par le modèle de régression logistique ainsi que leur significativité statistique. L'objectif est d'identifier les variables qui influencent le plus la probabilité de développer un diabète.

- **Variables fortement significatives (p < 0.001)** : Les variables suivantes sont statistiquement très significatives et ont une influence importante sur le risque de diabète :
  - **`HbA1c_level (Estimate = 2.318)`** : Un point supplémentaire d'HbA1c augmente fortement la probabilité d'être diabétique. Ceci est cohérent avec la littérature médicale; l'HbA1c est un marqueur direct du glucose sanguin sur plusieurs semaines.
  - **`Blood_glucose_level (Estimate = 0.03301)`** : L'augmentation du glucose sanguin accroît la probabilité de diabète. Même si son coefficient semble petit, la variable est mesurée en unités plus grandes (80 à 300), donc son effet cumulé est important.
  - **`Age (Estimate = 0.04698)`** : L'âge est également un facteur majeur; chaque année supplémentaire augmente la probabilité de diabète. Cela signifie que les risques augmentent progressivement au fil du temps.
  - **`Bmi (Estimate = 0.08526)`** : Un IMC élevé accroît le risque de diabète. Là aussi, l'effet cumulé est important (IMC allant de 10 à 95 dans le dataset)

- **`Hypertension (Estimate = 0.6814)`** : Les personnes hypertendues ont un risque nettement plus élevé de diabète. Cela s'aligne avec les études cliniques liant hypertension et troubles métaboliques.
- **`Heart_disease (Estimate = 0.6773)`** : Les individus ayant une maladie cardiaque ont également un risque beaucoup plus élevé de diabète. Le lien est logique car ces pathologies sont liées aux mêmes facteurs de risque (obésité, alimentation, mode de vie).
- **`genderMale (Estimate = 0.3056)`** : Être un homme est associé à une probabilité légèrement plus élevée de diabète par rapport aux femmes. L'effet est modéré mais significatif.
- **`smoking_historyNo Info (Estimate = -0.5674`**) : Cette catégorie a un effet négatif significatif. Cela signifie que les individus pour lesquels l'historique tabagique n'est pas documenté présentement une probabilité plus faible d'être diabétiques

- **Variables non significatives (p > 0.05)** : Certaines catégories de tabagisme ne montrent pas d'effet statistiquement significatif notamment : `smoking_historyever`, `smoking_historyformer`, `smoking_historynever`, `smoking_historynot current`, `genderOther` (extrêmement rare dans le dataset). Nous n'avons pas besoin de les interpréter mais nous les utiliserons car elles contribuent à la variabilité globale.

- **Lecture générale du modèle** : Le modèle identifie clairement quatre grands types de facteurs influençant le diabète :
  - **Facteurs biologiques** : HbA1c_level (plus fort prédicteur), blood_glucose_level
  - **Facteurs physiologiques** : age, bmi
  - **Facteurs cardio-vasculaires** : hypertension, heart_disease
  - **Facteur démographique** : genre (hommes davantage à risque que femmes)

En conclusion, L'analyse des coefficients du modèle montre que les variables `HbA1c_level`, `blood_glucose_level`, `age`, `bmi`, `hypertension` et `heart_disease` sont les facteurs les plus déterminants dans la prédiction du diabète. Ces résultats confirment que les indicateurs métaboliques et cardio-vasculaires jouent un rôle crucial dans la détection des individus à risque. Les catégories du tabagisme apparaissent moins informatives dans ce modèle. Cette interprétation permettra d'orienter les choix de modèles avancés lors des étapes ultérieures du projet.

4.3. Plan de modélisation pour la suite du projet

Les prochaines étapes du projet consisteront à tester plusieurs modèles de classification afin d'améliorer la détection du diabète, particulièrement dans un contexte de données déséquilibrées.

- **Gestion du déséquilibre :** Le jeu de données présente environ 9 % de cas positifs. Pour tenir compte de ce déséquilibre, trois approches seront testées :
  - **Sur-échantillonnage** de la classe minoritaire (ex. SMOTE)

- **Pondération des classes** dans certains modèles (ex. Random Forest, SVM)
- **Sous-échantillonnage** léger de la classe majoritaire

Dans un contexte de classes déséquilibrées, l'AUC-ROC est recommandée comme métrique principale notamment dans l'aide à la décision médicale *(Fawcett, 2006).* De même les Random Forest sont souvent considérés robustes aux données bruitées *(Breiman, 2001).*

- **Modèles envisagés :** d'autres algorithmes seront comparés;
  - **KNN (K-Nearest Neighbors)**
  - **Random Forest**
  - **SVM** (Support Vector Machine) : noyau linéaire et RBF

Chaque modèle sera entraîné avec validation croisée.

- **Prétraitement des données :** Les variables numériques seront normalisées ou standardisées lorsque nécessaire (KNN, SVM).
- **Métriques de performance :** Étant donné le déséquilibre des classes, les métriques suivantes guideront le choix du modèle final :
  - **AUC-ROC**
  - **Sensibilité (Recall)**
  - **F1-score**
  - **Matrice de confusion**

- **Sélection du modèle final :** Le meilleur modèle sera retenu sur la base :
  - de sa capacité à détecter les individus diabétiques
  - de sa performance globale (AUC, F1)
  - de sa stabilité entre les différents échantillons

5. Modèles avancés

Après avoir tester la régression logistique qui s'est avéré limite quant à sa capacité à détecter correctement les personnes atteintes de diabète, nous allons poursuivre avec des modèles d'algorithme d'apprentissage automatique plus avancés afin d'améliorer la capacité de détection du diabète, notamment dans un contexte de classes fortement déséquilibrées. Contrairement au modèle de référence, les modèles suivants sont capables de capturer des relations non linéaires d'intégrer des interactions complexes entre variables et d'offrir une meilleure sensibilité pour les cas positifs.
Les modèles à évaluer sont :

- **KNN (K-Nearest Neighbors)**
- **Random Forest**
- **SVM linéaire**
- **SVM à noyau RBF**

Pour chaque modèle, une normalisation des variables numériques sera appliquée lorsque cela est nécessaire (notamment pour KNN et SVM). L'évaluation sera réalisée à l'aide du jeu de

test, en utilisant les métriques suivantes : AUC-ROC, sensibilité (Recall), F1-score, Balanced Accuracy et matrice de confusion.

5.1. Modèle KNN

Le modèle KNN est une méthode d'apprentissage supervisé basée sur la similarité entre individus. Pour prédire la classe d'un nouvel individu, le modèle recherche les *k* observations les plus proches dans l'espace des caractéristiques, puis effectue un vote majoritaire. (Voir RMarkdown).

Interprétation des résultats :

Les premiers résultats **(performance générale)** montrent que :

- Le modèle KNN optimal utilise **k = 9 voisins.**
- Son **Accuracy** est élevée, ce qui est normal car la majorité des individus sont non diabétiques.
- Le **Kappa (0.63)** montre une performance correcte, mais inférieure à la régression logistique **(Kappa ~0.71).** On devra procéder à une vérification de la **sensibilité, F1 et AUC** pour juger la capacité du modèle à détecter les diabétiques (classe 1).

La suite des résultats **(détection des diabétiques et précision, qualité des prédictions, AUC-ROC)** montre :

- **Sensibilité = 0.505;** ce qui signifie que le modèle détecte 50.5% des diabétiques, soit une personne diabétique sur deux. Ce qui est loin des résultats de la régression logistique et très faible pour une modèle médical;
- **Spécificité = 0.996;** ce qui signifie que le modèle identifie très bien les non-diabétiques. Ce qui est normal car la majorité du dataset est en classe 0;
- **PPV (Pos Pred Value) = 92.38%.** Lorsqu'il prédit « diabétique », il a 92% de chances d'avoir raison; ce qui est très bon mais il prédit la classe 1 seulement dans les cas évidents;
- **Balanced Accuracy = 0.75**; Ce score est correct, mais nettement inférieur à la régression logistique (~ 0.82). Ce qui confirme que KNN gère mal le déséquilibre des classes.
- **AUC = 0.9153.** Le modèle a une excellente capacité globale de discrimination entre les diabétiques et non diabétiques. Mais l'AUC élevé ne compense pas la sensibilité trop faible qui est un élément important de la prédiction.

En conclusion le modèle KNN avec k =9 obtient un AUC très élevé, ce qui montre qu'il distingue bien les classes en termes de probabilité. Cependant, sa sensibilité est insuffisante pour un contexte médical, quasiment 50% des diabétiques non détectés. Ainsi le modèle KNN qui a des performances globales inférieures à la régression logistique surtout en terme de sensibilité, ne sera pas retenu comme meilleur modèle.

5.2. Modèle Random Forest

Ce modèle est un ensemble d'arbres de décision. Il construit plusieurs arbres et la classe finale est décidée par vote majoritaire. (Voir RMarkdown).

Interprétation des résultats :

Les performances générales montrent :

- **Accuracy = 97.17%.** Ce qui est très élevé, mais à interpréter avec prudence à cause du déséquilibre des classes.
- **Kappa = 0.7946** ; C'est un excellent niveau d'accord supérieur à celui du KNN = 0.63 et meilleur que la régression logistique (0.72). Cela montre une progression réelle de la qualité prédictive.

Les paramètres de détection des diabétiques, de précision et de qualité de prédictions montrent :

- **Sensibilité = 0.6796.** En d'autres termes le modèle détecte 67.9% des diabétiques; ce qui est mieux que la régression logistique (64%) et le modèle KNN (50%). Le modèle Random Forest détecte plus de patients à risque.
- **Spécificité = 1.000 ;** Aucun faux positif; le modèle prédit correctement et toujours lorsqu'il dit diabétique.
- **PPV (Pos Pred Value) = 100%.** Ce qui signifie que lorsque le modèle prédit diabète, il ne se trompe jamais, donc le modèle est extrêmement strict dans ses prédictions
- **NPV (Neg Pred Value) = 96.99%;** les prédictions de classe 0 (non diabétiques) sont aussi très fiables
- **Balanced Accuracy = 0.8398.** Il est correct et légèrement supérieur à la régression logistique
- **AUC = 0.9216 :** excellente capacité à distinguer les diabétiques des non diabétiques. Il est plus élevé que dans la régression logistique et le modèle KNN.

En conclusion, le modèle Random Forest est très performant sur ce dataset, avec une bonne sensibilité à détecter les personnes diabétiques, une prédiction sans faute des cas de diabètes (PPV = 100%) mais surtout une excellente capacité à distinguer les diabétiques des non diabétiques (AUC = 0.9216). Actuellement il est reste le modèle parfait pour notre étude par rapport à la régression logistique et le KNN, mais pour plus d'assurance nous allons procéder au test avec les autres modèles cités.

5.3. Le modèle SVM linéaire

Le SVM linéaire cherche à séparer les deux classes (0 et 1) avec une frontière droite (linéaire). Il maximise la marge entre les deux groupes, ce qui donne un modèle simple et efficace. (Voir RMarkdown).

Interprétation des résultats :

Les performances générales montrent :

- **Accuracy = 96.09%,** comme pour le modèle Random Forest, il est élevé mais du fait du déséquilibre il faut l'interpréter avec prudence;
- **Kappa = 0.7223**; très bon niveau d'accord, performance en dessous du modèle Random Forest et supérieure aux modèles Régression logistique et KNN;

Les données de détection des diabétiques, de précision, de qualité de prédiction et AUC donnent les informations suivantes :

- **Sensibilité = 0.6411**; le modèle détecte parfaitement 64.1% des diabétiques, inférieur au modèle Random Forest (~ 68%)
- **Spécificité = 0.9918**. Le modèle identifie très bien les non-diabétiques
- **PPV = 88.3%;** lorsqu'il prédit diabète, il a 88% de chances d'avoir raison. C'est certes élevé mais inférieur aux 100% de Random Forest.
- Balanced Accuracy = 0.816; bon score mais encore inférieur à celui de Random Forest (~ 0.84)
- **AUC = 0.9624.** Ceci une excellente capacité à distinguer les diabétiques des non-diabétiques. C'est la plus grande AUC de tous les modèles testés à juste 4 points de Random Forest. Cela signifie aussi qu'en terme de probabilités, le SVM linéaire discrimine extrêmement bien les deux classes.

En conclusion, le SVM Linéaire a un bon AUC par rapport aux autres modèles et surtout le Random Forest, sa sensibilité et sa spécificité sont également bonnes; mais dans l'ensemble sa performance globale est inférieure à Random Forest notamment pour détecter les diabétiques, prédire les positifs.
Ce modèle sera intéressant pour la comparaison finale, même si Random Forest reste devant en termes de détection et de prédiction des diabétiques.

Une fois les différents modèles testés, il nous faut procéder aux comparaisons et aux tracés des courbes.

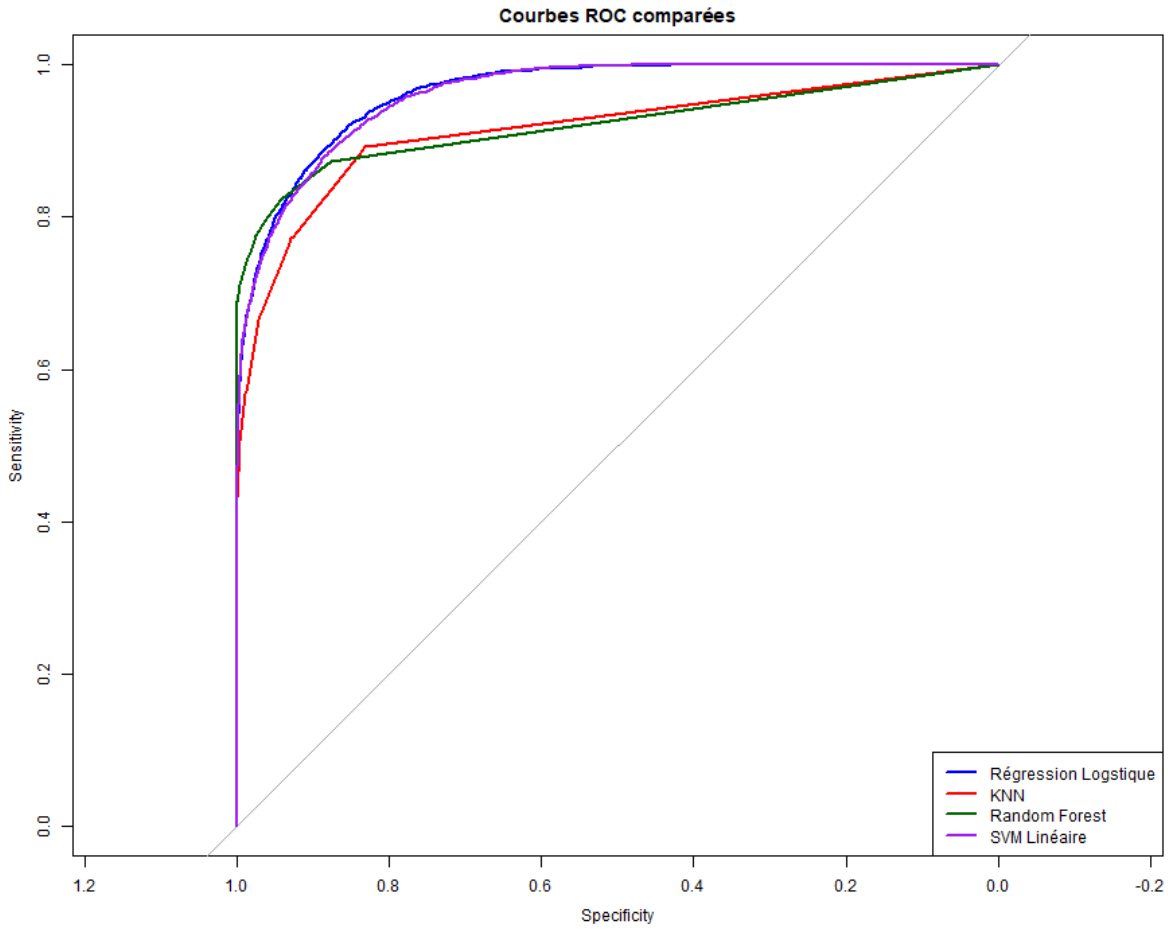
6. Comparaisons des modèles

6.1. Tableau comparatif des modèles

Pour mieux illustrer les différentes variables de chaque modèle, nous allons créer un dataframe (Voir RMarkdown).

6.2. Tracé des courbes ROC des modèles

Le tracé des courbes ROC permet de mieux observer la superposition des modèles (Voir RMarkdown).



<u>Interprétation du graphique :</u> Le tracé de la courbe permet de visualiser les performances globales des quatre modèles testés. Plus une courbe est approchée du coin supérieur gauche, meilleure est la capacité du modèle à distinguer les individus diabétiques des non-diabétiques. On peut ainsi observer :

- **La régression logistique (bleu) et le SVM linéaire (violet)** affichent les meilleures performances globales. Leurs courbes sont les plus proches du coin supérieur gauche, ce qui correspond à leurs AUC très élevées (0.96);
- **Le Random Forest (vert)** obtient également une courbe très performante, légèrement en dessous des deux précédents, ce qui est cohérent avec son AUC de 0.92. Sa forme assez régulière confirme sa bonne stabilité;
- **Le KNN (rouge)** présente la courbe la moins performante des quatre modèles. Sa ROC est plus éloignée du coin supérieur gauche, reflétant une capacité de séparation plus limitée (AUC = 0.91).

De manière générale, le graphique montre clairement que les modèles **SVM linéaire, régression logistique et Random Forest** surpassent le KNN en termes de discrimination, avec un avantage visuel pour le SVM linéaire et la régression logistique, dont les courbes dominent la figure.

6.3. Discussion critique

La comparaison des modèles montre des performances contrastées selon les métriques. Le KNN est le moins performant, avec une sensibilité faible (50%) et une balanced accuracy limitée, ce qui confirme qu'il détecte mal les cas de diabète.

La régression logistique et le SVM linéaire obtiennent d'excellents scores de discrimination (AUC = 0.96), ce qui est confirmé visuellement par leurs courbes ROC proches du coin supérieur gauche. Cependant, leurs sensibilités restent modestes (64%), avec un léger avantage pour la régression logistique.

Le **Random Forest** se démarque comme le modèle le plus équilibré; il présente la meilleure sensibilité, une balanced accuracy élevé et un Kappa supérieur. Sa courbe ROC est également solide, bien que légèrement inférieure aux modèles linéaires en termes d'AUC.

Globalement, **Random Forest** apparaît comme le meilleur modèle pour la détection des individus diabétiques, tandis que la **régression logistique et le SVM linéaire** excellent surtout en capacité de discrimination globale. A travers tous ses éléments nous pouvons maintenant faire le choix de notre modèle final approprié.

7. Sélection du modèle final

Le choix du modèle final repose sur une appréciation entre performance globale et capacité à détecter correctement les individus diabétiques, qui représentent la classe d'intérêt dans ce contexte.

Parmi les différents modèles testés et évalués, Random Forest se démarque comme le plus performant pour la détection du diabète. Il présente la meilleure sensibilité, la balanced accuracy la plus élevée et un Kappa supérieur à l'ensemble des autres modèles. De plus, il ne produit aucun faux positif, ce qui renforce la fiabilité de ses performances positives (PPV =100%).

Même si la régression logistique et le SVM linéaire obtiennent des AUC légèrement plus élevées, ils présentent tous deux une sensibilité plus faible, ce qui se traduit par davantage de faux négatifs. Or dans un contexte de dépistage du diabète, la priorité est de maximiser la détection des cas positifs, même si cela implique une légère baisse de performance discriminante.

Par conséquent, le modèle ***RANDOM FOREST*** est retenu comme modèle final, du fait qu'il offre un meilleur équilibre entre la détection des diabétiques, la performance globale et robustesse prédictive.

De plus selon **Breiman (2001)**, le **Random Forest (Forêts aléatoires)** présente une excellente performance sur les données tabulaires, avec une forte capacité de généralisation et une

robustesse au surajustement. Ces propriétés expliquent les bons résultats obtenus dans ce projet, notamment en termes de sensibilité et balanced accuracy.

Comme l'on souligné **Cutler et al. (2012)** les Random Forest se distinguent par leur précision élevée et leur résistance au surapprentissage, ce qui explique leur performance solide dans ce travail.

8. Lien vers le dépôt GitHub

L'ensemble des fichiers utilisés dans ce travail (RMarkdown, HTML, Modèle final, jeux de données et rapport Word) est disponible dans le dépôt GitHub suivant :

https://github.com/Steve8965/Projet_SciencesDeDonnees_Diabetes

Ce dépôt contient :

- Le dossier **Projet Final** contenant le fichier **SCI1402_Plan du Projet.pdf** et le fichier **ProjetFinal_OrlySteve_NgayapTchouamdjou.pdf** qui est le projet final soumis à la TÉLUQ
- Le **Modèle_Final_Random_Forest.rds** qui correspond au modèle final choisi
- Le fichier **SCI1402_Projet en sciences de données_NgayapTchouamdjou_OrlySteve.Rmd** contenant les codes R complet
- Le fichier **SCI1402_Projet en sciences de données_NgayapTchouamdjou_OrlySteve.html** qui est le rapport exécuté dans RMarkdown.
- Le fichier **Data** contenant les deux jeux de données (données initiales et données nettoyées)
- Le fichier **Images_ProjetSciencesDeDonnees** contenant les images générées par exécution des codes et insérés dans le rapport.

Références :

World Health Organization. (2023). *Diabetes fact sheet.*

American Diabetes Association. (2023). *Standards of medical care in diabetes.*

Breiman, L. (2001). Random forests. *Machine Learning, 45, 5–32.*

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27(8), 861–874*
International Diabetes Federation. (2023). *IDF Diabetes Atlas.*

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2012). *Random forests for classification in ecology*. **Ecology, 88**(11), 2783–2792.