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

Travail noté 1 : Plan de projet documenté

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 : 03 novembre 2025

PLAN DE PROJET

TYPE DE PROBLEME ET ENJEUX DU PROJET

Le diabète est une maladie chronique en forte croissance mondiale, représentant un enjeu majeur de santé publique. La détection précoce des personnes à risque permettrait d'améliorer la prévention, de réduire les complications à long terme et d'optimiser les coûts du système de santé. Le projet s'inscrit dans une problématique de classification prédictive : il s'agit d'identifier, à partir de variables biométriques et comportementales, les individus les plus susceptibles de développer un diabète.

L'enjeu principal est de démontrer comment les méthodes d'apprentissage automatique peuvent appuyer la prise de décision en santé publique, en fournissant des modèles capables d'estimer le risque individuel à partir de données cliniques et sociodémographiques. Ce type de solution s'inscrit dans la mouvance actuelle de la médecine préventive et personnalisée, tout en illustrant les applications concrètes de la science des données en santé.

ANALYSE DU CONTEXTE

Selon la Fédération internationale du diabète (IDF, 2024), plus de 589 millions d'adultes vivent avec cette maladie, et ce chiffre pourrait atteindre 780 millions d'ici 2045. Environ 3,4 millions de décès lui sont attribués chaque année. Au Canada, selon Diabetes Canada (2024), environ 10% de la population en est atteinte, un taux qui grimpe à 15% si l'on inclut les cas non diagnostiqués. Les coûts associés au traitement et aux complications du diabète augmentent considérablement chaque année, soulignant la nécessité de solutions prédictives et préventives.

Le jeu de données choisi, *« Diabetes Prediction Dataset »* disponible sur www.Kaggle.com , regroupe plusieurs variables (âge, indice de masse corporelle, tension artérielle, antécédents familiaux, niveau d'activité physique, etc.) et une variable cible indiquant la présence ou non de diabète. Ce contexte offre une base idéale pour appliquer des techniques de prétraitements, modélisation, validation et évaluation dans une approche scientifique complète.

DEFINITION DES OBJECTIFS ET HYPOTHESES

Objectif général

- Développer un modèle d'apprentissage automatique capable de prédire avec précision le risque de diabète à partir de caractéristiques biométriques et comportementales.

Objectifs spécifiques :

- Explorer et nettoyer les données afin d'assurer leur qualité et leur cohérence;
- Identifier les variables les plus corrélées au risque de diabète;

- Tester et comparer plusieurs modèles de classification (régression logistique, KNN, forêts aléatoires, SVM, etc.);
- Évaluer la performance des modèles à l'aide de métriques appropriées (AUC-ROC, précision, rappel, F1-score);
- Interpréter les résultats afin de mettre en évidence les facteurs les plus déterminants dans le risque de diabète.

Hypothèses principales :

- Il existe des corrélations significatives entre certaines variables (comme le BMI, la pression artérielle et les antécédents familiaux) et la probabilité de diabète.
- Les modèles basés sur des approches non linéaires (forêts aléatoires, SVM) offriront une meilleure performance prédictive qu'une régression logistique simple.
- L'évaluation à l'aide de l'AUC-ROC permettra de mesurer efficacement la capacité du modèle à distinguer entre individus diabétiques et non diabétiques.

SOMMAIRE DE LA MÉTHODOLOGIE

Ce projet vise à analyser un jeu de données médicales comportant diverses variables biologiques et comportementales afin de construire un modèle prédictif du risque de diabète. L'étude sera réalisée à l'aide du jeu de données *Diabetes Prediction Dataset* disponible sur Kaggle.

Le projet suivra les étapes clés d'un processus complet de sciences des données :

1. **Collecte et préparation des données :** importation du jeu de données *Diabetes Prediction Dataset* dans RStudio, inspection des types de variables continues et encodage des catégorielles.
2. **Analyse exploratoire :** visualisation statistique et graphique à l'aide des bibliothèques ***tidyverse, ggplot2 et corrplot*** pour explorer les distributions, corrélations et tendances des variables explicatives.
3. **Modélisation :** mise en œuvre et comparaison de plusieurs algorithmes de classification supervisée (régression logistique, KNN, Randon Forest, SVM) à l'aide des bibliothèques ***caret et e1071***.
4. **Évaluation :** comparaison des modèles selon les métriques classiques (précision, rappel, F1-score, AUC-ROC) et visualisation des courbes ROC.
5. **Interprétation et communication des résultats :** présentation des variables les plus influentes sur le risque de diabète, accompagnée de visualisations synthétiques et d'un rapport final généré sous RMarkdown.