الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

University Setif 1– FERHAT ABBAS

Faculty of Natural and Life Sciences

Department of Biochemistry

جامعة سطيف1-فرحات عباس

كلية علوم الطبيعة والحياة

DERARTEMENT OF BIOCHEMISTRY

N°……/SNV/2025

# Memory

Presented by

**Choubane Sara**

**Djerboua Mebarka Besma**

For the fulfillment of the requirements for the degree of

## Master

**Field: Biological Sciences**

**Special field: Applied Biochemistry**

## TOPIC

# Developing a Machine Learning-Driven Clinical Decision Support System for Early Detection and Management of Diabetes Mellitus

**Presented in 21/06/2026**

**JURY:**

| | | |
|---|---|---|
| **Chair:** | **Pr Ouksel Louiza** | **Pr UFA Setif 1** |
| **Supervisor:** | **Dr Aouina Nabila Yasmina** | **MCA UFA Setif 1** |
| **Examiner:** | **Dr Khaznadar Mouna** | **MCA UFA Setif 1** |

# List of contents


**Acknwoledgements** **i**

**Abstracts** v

**Liste of abréviation** **viii**

**Liste of figures** **x**

**Liste of tables** **xiii**

**General introduction** **1**

**Chapter 1 : Background & Literature Review**

**1.1.** Introduction **3**

**1.2.** Diabetes Mellitus **3**

**1.2.1.** Types of diabetes **3**

**1.2.2.** Causes of diabetes **4**

**1.2.3.** Risk factors **5**

**1.2.4.** Symptoms and complications **6**

**1.3.** Traditional diagnostic methods for diabetes **6**

**1.3.1.** Hemoglobin A1c (HbA1c) Test **6**

**1.3.2.** Oral Glucose Tolerance Test (OGTT) **7**

**1.3.3.** Fasting Plasma Glucose (FPG) **8**

**1.4.** Artificial Intelligence and Machine Learning in Diabetes Diagnosis **8**

**1.4.1.** Artificial Intelligence in Healthcare **8**

**1.4.2.** Machine learning techniques for diabetes prediction **8**

**1.4.3.** Clinical Decision Support Systems **9**

**1.4.4.** Importance of early prediction models **9**

**1.5.5.** Challenges in ML-based diabetes prediction **9**

**1.6.** Conclusion **9**

## Chapter 2 : Materials and methods


**2.1.** Introduction **11**

**2.2** Machine Learning models **11**

**2.2.1.** Logistic Regression **11**

**2.2.2.** Decision trees **12**

**2.2.3.** Support Vector Machine **14**

**2.2.4.** Random Forest **15**

**2.2.5.** Extreme Gradient Boosting (XGBoost) **17**

**2.3.** Model evaluation metrics **18**

**2.3.1.** Accuracy **18**

**2.3.2.** Precision **18**

**2.3.3.** F1-score **19**

**2.3.4.** ROC-AUC **19**

**2.3.5.** Recall **20**

**2.4.** Conclusion **21**


## Chapter 3: Exploratory Data Analysis and Preprocessing


**3.1.** Introduction **22**

**3.2.** Exploratory Data Analysis (EDA) **22**

**3.2.1.** The dataset description **22**

**3.2.2.** Descriptive statistics **23**

**3.2.3.** Data quality checks **25**

**3.2.4.** Univariate analysis **26**

**3.2.5.** Class imbalance analysis **27**

**3.2.6.** Bivariate analysis **28**

**3.3.** Data Preprocessing **32**

**3.3.1.** Data cleaning **32**

**3.3.2.** Train–Test Split **32**

**3.3.3.** Encoding categorical variables **33**

**3.3.4.** Feature scaling (Standardization) **33**

**3.3.5.** Handling class imbalance **33**

**3.4.** Conclusion **33**

## Chapter 4: Exploratory Data Analysis and Preprocessing

**4.1.** Introduction **34**

**4.2.** Machine Learning modeling **34**

**4.2.1.** Logistic Regression **34**

**4.2.2.** Random Forest **35**

**4.2.3.** XGBoost **36**

**4.2.4.** Decision Tree **37**

**4.2.5.** Support Vector Machine **39**

**4.2.6.** Multi-model ROC curve comparison **40**

**4.2.7.** SHAP Analysis **41**

**4.3.** Conclusion **42**

**General conclusion** **44**

**References**

**الملخص**

تتناول هذه الدراسة استخدام تقنيات التعلم الآلي للتنبؤ بالتشخيص المبكر لمرض السكري بالاعتماد على مجموعة بيانات كبيرة تضم حوالي 100,000 سجلٍ للمرضى وتسع خصائص ديموغرافية وسريرية. وقد تم تطبيق سلسلة شاملة من خطوات المعالجة المسبقة لتحسين جودة البيانات وأداء النماذج، شملت تنظيف البيانات، وترميز المتغيرات، والتطبيع (Normalization)، ومعالجة مشكلة عدم توازن الفئات. تم تدريب وتقييم خمسة خوارزميات للتعلم الآلي الخاضع للإشراف، وهي: (الانحدار اللوجستي، وشجرة القرار، والغابة العشوائية، وآلة المتجهات الداعمةSVM ، وخوارزميةXGBoost) ، وذلك باستخدام أسلوب التحقق المتقاطع وعدة مؤشرات لتقييم الأداء .أظهرت النتائج أن جميع النماذج المطورة تمتلك قدرات تنبؤية عالية، حيث تجاوزت دقة التصنيف 95%. ومن بين هذه النماذج، حققت خوارزمية XGBoost أفضل أداء، بدقة بلغت 97.21%، وقيمة بلغت 99.24%، ومساحة تحت منحنى ROC (ROC-AUC) بلغت 97.67%. ولتعزيز قابلية تفسير النماذج التنبؤية، تم استخدام طريقة SHAP لتحديد مساهمة كل متغير في تنبؤات النموذج. وأظهرت نتائج التحليل أن مستوى الهيموغلوبين السكري (HbA1c) ومستوى سكر الدم كانا الأكثر تأثيرًا في التنبؤ بمرض السكري، يليهما العمر ومؤشر كتلة الجسم، مما يؤكد الدور الحاسم لهذه المتغيرات في تقييم خطر الإصابة بالمرض.

تُبرز هذه النتائج فعالية تقنيات التعلم الآلي، وخاصة أساليب التعلم التجميعي مثل XGBoost ، كوسيلة واعدة للكشف المبكر عن مرض السكري ودعم أنظمة اتخاذ القرار السريري.

**الكلمات المفتاحية :**داء السكري، التنبؤ المبكر، التعلم الآلي، XGBoost، تحليلSHAP ، أنظمة دعم القرار السريري.

**Abstract**

This study examines the use of machine learning techniques to predict early diagnosis of diabetes mellitus using a large dataset including approximately 100,000 patient records and nine demographic and clinical characteristics. A comprehensive preprocessing pipeline was implemented to enhance data quality and model performance: data cleaning, feature encoding, normalization, and class imbalance correction. Five supervised machine learning algorithms (logistic regression, decision tree, random forest, support vector machine, and XGBoost) were trained and evaluated through cross-validation and using several performance indicators. The results showed that all the developed models possessed strong predictive capabilities, with classification rates exceeding 95%. XGBoost, among them, provided the best performance with an accuracy of 97.21%, a correctness of 99.24%, and a ROC-AUC score of 97.67%. To improve the interpretability of the predictive models, the SHAP method was used to determine the contribution of each variable to the model's predictions. The analysis showed that HbA1c and blood glucose levels were the most influential variables, followed by age and body mass index (BMI), highlighting their crucial role in assessing diabetes risk.These results demonstrate that machine learning, particularly holistic approaches like XGBoost, can be an effective means of early diabetes screening.

**Keywords:** Diabetes Mellitus, Early prediction, ML, XGBoost, SHAP analysis, Clinical Decision Support Systems.

**Résumé**

Cette étude examine l'utilisation des techniques d'apprentissage automatique pour le diagnostic précoce du diabète sucré à partir d'un vaste jeu de données comprenant environ 100 000 dossiers de patients et neuf caractéristiques démographiques et cliniques. Une chaîne complète de prétraitement a été mise en œuvre afin d'améliorer la qualité des données et les performances des modèles, incluant le nettoyage des données, l'encodage des variables, la normalisation et la correction du déséquilibre des classes. Cinq algorithmes d'apprentissage supervisé — la régression logistique, l'arbre de décision, la forêt aléatoire, la machine à vecteurs de support (SVM) et XGBoost — ont été entraînés et évalués à l'aide de la validation croisée et de plusieurs métriques de performance. Les résultats ont montré que tous les modèles développés présentaient d'excellentes capacités prédictives, avec des taux de classification supérieurs à 95 %. Parmi eux, XGBoost a obtenu les meilleures performances, avec une exactitude de 97,21 %, une précision de 99,24 % et une aire sous la courbe ROC (ROC-AUC) de 97,67 %. Afin d'améliorer l'interprétabilité des modèles prédictifs, la méthode SHAP (SHapley Additive exPlanations) a été utilisée pour évaluer la contribution de chaque variable aux prédictions. L'analyse a révélé que le taux d'HbA1c et la glycémie constituaient les facteurs les plus influents, suivis de l'âge et de l'indice de masse corporelle (IMC), soulignant leur rôle déterminant dans l'évaluation du risque de diabète. Ces résultats démontrent que l'apprentissage automatique, en particulier les méthodes d'ensemble telles que XGBoost, constitue un outil efficace pour le dépistage précoce du diabète et peut contribuer au développement de systèmes d'aide à la décision clinique plus précis et plus fiables.

**Mots-clés :** Diabète sucré, prédiction précoce, apprentissage automatique, XGBoost, analyse SHAP, systèmes d'aide à la décision clinique.

# General introduction

Diabetes mellitus is a chronic metabolic disorder characterized by persistent hyperglycemia resulting from impaired insulin secretion, insulin action, or both. It represents a major global public health challenge, with a continuously increasing prevalence worldwide. According to recent estimates, approximately 463 million adults aged between 20 and 79 years were living with diabetes, and the disease, together with its associated complications, caused nearly 4.2 million deaths in 2019 (Sun et al., 2022).

Early detection of diabetes is essential for reducing disease burden and preventing severe long-term complications. Numerous studies have demonstrated that screening programs, HbA1c testing, and artificial intelligence-based predictive approaches can identify high-risk individuals before irreversible organ damage occurs (Yesupatham et al., 2025). However, conventional diagnostic methods still present several limitations. The Oral Glucose Tolerance Test (OGTT), although widely used, is time-consuming and relatively expensive due to the prolonged waiting period required during testing. Similarly, HbA1c measurements may produce unreliable results in patients suffering from specific medical conditions, such as liver disease, which can lead to diagnostic inaccuracies (Lahousen et al., 2004).

In recent years, machine-learning techniques have emerged as powerful tools for addressing complex healthcare challenges. Due to their ability to process large volumes of medical data and identify hidden relationships among variables, ML models have significantly improved the accuracy, efficiency, and reliability of predictive systems (Al-Nafjan et al., 2025). These approaches have been successfully applied in numerous medical domains, including cancer diagnosis, neurological disorder detection, medical imaging, clinical chemistry, and wearable healthcare technologies (Erickson et al., 2017). Consequently, ML has become an important component of modern clinical decision-support systems and continues to contribute to advancements in medical research and healthcare innovation (Çubukçu et al., 2024).

Despite the availability of traditional diagnostic techniques, diabetes remains one of the leading causes of morbidity worldwide because of delayed diagnosis and the progression of severe complications. Therefore, developing accurate and efficient machine learning models for early diabetes prediction has become increasingly important (Sun et al., 2022).

The primary objective of this study is to develop and evaluate machine-learning models for diabetes prediction. The work involves dataset analysis and preprocessing, implementation of different machine learning algorithms, and comparative evaluation of their predictive

performance using appropriate assessment metrics. The ultimate aim is to identify the most effective predictive model for supporting the early detection of diabetes.

This manuscript is organized into four chapters. Chapter I provides a general overview of diabetes mellitus, including its types, causes, symptoms, risk factors, complications and the use of Machine learning in diabetes diagnosis. Chapter II presents the materials and methods employed in this study, with emphasis on machine learning algorithms and evaluation metrics. Chapter III focuses on exploratory data analysis and data preprocessing procedures. Finally, Chapter IV presents the experimental results, evaluates the performance of the developed models, and discusses the main findings of the study.