وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم البرمجيات


# التشخيص المبكر لمرض السكري باستخدام التعلم الآلي

رسالة مقدمة
إلى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دبلوم عالي في
البرمجيات

من قبل
ايهاب طارق الياس خلو

بإشراف
م.د منى محمد طاهر جوهر

١٤٤٥ هـ ٢٠٢٤م

## الخلاصة

مرض السكري (Diabetes Mellitus - DM) هو أحد الاضطرابات الأيضية التي تؤدي إلى ارتفاع مستويات السكر في الدم بشكل غير مناسب. وهو مرض مزمن شائع ويشكل خطرا كبيرا على صحة الإنسان. ومن سمات مرض السكري ارتفاع نسبة الجلوكوز في الدم عن المستوى الطبيعي.

ويعتبر التعلم الالي هو احد فروع الذكاء الاصطناعي ، يقوم بالتركيز على انشاء انظمة تقوم بتعلم البيانات واكتساب المعرفة لتحسين ادائها بشكل تلقائي، يعتمد التعلم الالي على الخوارزميات والنماذج التي تسمح للانظمة بتحليل البيانات واكتساب الخبرة واتخاذ القرارات مما يمكنها من التكيف مع المهام وتحسين ادائها مع مرور الوقت . ساعدت خوارزميات التعلم الآلي العاملين في مجال الصحة (بما في ذلك الأطباء) في معالجة المشكلات الطبية وتحليلها وتشخيصها. يمكن للتعلم الآلي أن يساعد الأشخاص على إصدار حكم أولي حول مرض السكري وفقًا لبيانات الفحص البدني اليومي، ويمكن أن يكون بمثابة مرجع للأطباء.

في هذه الرسالة تم استخدام خوارزميات التعلم الالي للكشف عن مرض السكري . تتضمن الخوارزميات المستخدمة كل من الانحدار اللوجستي Logistic Regression (LR) ، شجرة القرار Decision Tree (DT) ، الغابة العشوائية Random Forest (RF) ، ألة المتجه الداعم Support vector machine ( SVM ) ، الجار الاقرب K-Nearest Neighbor (KNN) و بايز البسيط Naive Bayes ( NB ) . تم جمع البيانات المستخدمة في هذا الرسالة من مجموعة بيانات Kaggle العالمية المعتمدة من قبل الباحثين. وقد خضعت البيانات لأنواع مختلفة من المعالجة المسبقة لتحسين اداء النموذج. تم تقييم الخوارزميات المستخدمة باستعمال مجموعة من المقاييس مثل الدقة Accuracy ،معدل الاستدعاء Recall، درجة F1 (F1-score) ، الضبط Precision واخيرا مصفوفة الارتباك Confusion Metrics وكانت اعلى دقة تصنيف ثنائي تم الحصول عليها عند تنفيذ الخوارزميات الستة هي لخوارزمية شجرة القرار ٩٨,٦٦٪ تليها الة المتجه الداعم ٨٥,٣٣٪ ثم خوارزمية الجار الاقرب ٨١٪ وخوارزمية بايز البسيط ٨٠٪ ثم خوارزمية الانحدار اللوجستي ٧٧,٦٦٪ واخيرا خوارزمية الغابة العشوائية ٧٦,٦٦٪ .

Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Software


# Early Diagnosis of Diabetes Using Machine Learning

**Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Higher diploma
in
Software**

**By**

**Ihab Tareq Elias Khalow**

**Supervised by**

**Dr. Muna Mohammed Taher Jawher**

---

**2024 A.D.** **1445 A.H.**

# Abstract

Diabetes Mellitus - DM is a metabolic disorder that leads to inappropriately high blood sugar levels. It is a common chronic disease and poses a major threat to human health. One of the characteristics of diabetes is that the blood glucose level is higher than the normal level.

Machine learning is one of the branches of artificial intelligence that focuses on creating systems that learn data and acquire knowledge to improve their performance automatically. Machine learning depends on algorithms and models that allow systems to analyze data, gain experience, and make decisions, enabling them to adapt to tasks and improve their performance over time. Machine learning algorithms have helped health professionals (including doctors) treat, analyze and diagnose medical problems, as well as detect disease patterns and other patient data. Machine learning can help people make an initial judgment about diabetes according to daily physical examination data, and can serve as a reference for doctors.

in this thesis, Machine learning algorithms were used to detect diabetes. The algorithms used in the research include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). And Naive Bayes (NB). The data used in this research was collected from the global Kaggle dataset approved by the researchers. The data has been subjected to different types of pre-processing to improve the model's performance. The algorithms used were evaluated using a set of metrics such as Accuracy, Recall Rate, F1-score, Precision, and finally Confusion Metrics. The highest binary classification accuracy obtained when implementing the six algorithms was for the decision tree algorithm, 98.66%, followed by 98.66%. The support vector machine 85.33%, then the nearest neighbor algorithm 81%, the simple Bayes algorithm 80%, then the logistic regression algorithm 77.66%, and finally the random forest algorithm 76.66%.