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


# إنشاء أداة قائمة على التعلم الآلي لتقييم مخاطر الأمن السيبراني

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

من قبل
عمر ابراهيم شيت المعاضيدي

بإشراف
ا.د. لهيب محمد ابراهيم الزبيدي

1444 هـ 2023 م

# الخلاصة

على مدى السنين الماضية، ازداد عدد الهجمات السيبرانية في جميع أنحاء العالم ، والتي أثرت بشكل متزايد على الشبكات والأنظمة والشركات ، وهذه الاضرار ترتفع كل عام . في الآونة الأخيرة ، طور الباحثون والشركات أطر عمل لتقييم مخاطر الأمن السيبراني من أجل تحديد وتقدير وترتيب أولويات المخاطر السيبرانية وتقليل تأثيرها. ومع ذلك ، غالبًا ما تكافح النهج التقليدية للعثور على مؤشرات للمخاطر السيبرانية غير المتوقعة، مما يحد من القدرة على إجراء تقييمات دقيقة للمخاطر.

من أجل تقييم مخاطر الأمن السيبراني في هذه الرسالة تم توليد نوعين من البيانات الخاصة بالمخاطر ((data risk 3 level (dr3l)، data risk 5 level (dr5l)) كل نوع تم توليد 4 مجاميع، وتم ايضا تجميع بيانات من موقع (vuldb) ومن ثم بناء أداة لتقييم هذه المخاطر باستخدام خوارزميات (آلة تعزيز التدرج الخفيف Light Gradient Boosting (LGBM Machine، الانحدار اللوجستي(Logistic Regression (LR، الغابة العشوائية Random Forest (RF)، تعزيز التدرج الاقصى(eXtreme Gradient (XGBoost Boosting، CatBoost، الشبكة العصبية الاصطناعية المدرك Multi-Layer Perceptron(MLP)، النموذج المهجن بأستخدام المكدس والذي يتكون من خوارزميتي LGBMو LR) .

تم إجراء عملية تطبيع للبيانات وبعد ذلك تقييم المخاطر بأستخدام نماذج الأداة وحساب مقاييس التقييم (Accuracy، Precession، Recall، F1-Score، Confusion Matrix) بالاعتماد على التقييم المقدر والتقييم الحقيقي والمقارنة بين النماذج وإبراز الافضل من بين النماذج لكل مجموعة من مجاميع البيانات.

تم اجراء مقارنة بين نماذج الأداة بأستخدام مقاييس التقييم لكل مجموعة بيانات لتوضيح النموذج الذي يعطي افضل اداء ففي مجموعة بيانات dr3l افضل اداء كان للشبكة العصبية الاصطناعية MLP إذ كانت الدقة Accuracy 99.84 % ، اما في مجموعة بيانات dr5l افضل اداء كان لخوارزمية آلة تعزيز التدرج الخفيف LGBM إذ كانت الدقة Accuracy 99.74 %، اما مجموعة بيانات data vulnerability افضل اداء كان لخوارزمية CatBoost إذ كانت الدقة Accuracy 99.38 %.

كما تم مقارنة اداء نماذج الأداة مع عمل سابق الذي استخدم نفس المقاييس مع نفس مجموعة البيانات ذات ال 50,000 عينة الذي تضمن اربع نماذج (شجرة القرار (DT) Decision tree، الجار الاقرب K-Nearest Neighbors (K-NN)، آلة المتجه الداعمة

(MLP المدرك الاصطناعية العصبية الشبكة ،Support vector machine (SVM
واظهرت النتائج تفوق نماذج الأداة على نماذج العمل السابق حسب مقاييس التقييم (Accuracy
، Precession ،Recall ،F1-Score ،Confusion Matrix).

University of Mosul
College of Computer Sciences
And Mathematics
Department of Software
Engneeing

# Construct a Tool based on Machine Learning for Cyber security Risk Assessment

**A 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 Master of Science
in
Software**

**By**
**Omar Ibrahim Sheet Almaathede**

**Supervised By**
**Prof.Dr. Laheeb Mohammed Ibrahim Al-Zobaidy**

**2023 A.D.** **1444 A.H.**

## Abstract

Over the past years, the number of cyber attacks has increased all over the world, which has increasingly affected networks, systems and companies, and this damage is rising every year. Recently, researchers and companies have developed cybersecurity risk assessment frameworks in order to identify, quantify, and prioritize cyber risks and minimize their impact. However, traditional approaches often struggle to find indicators of unexpected cyber risks, which limits the ability to make accurate risk assessments.

In order to assess the cybersecurity risks in this message, two types of risk data were generated: data risk 3 level (dr3l), data risk 5 level (dr5l), each type generated 4 totals, and data was also collected from (vuldb) and from Then build a tool to assess these risks using algorithms Light Gradient Boosting Machine(LGBM), Logistic Regresion (LR), Random Forest (RF) , eXtreme Gradient Boosting (XGBoost), CatBoost, Multi-Layer Percetron (MLP), a hybrid model using the stack which consists of LGBM and LR algorithms.

A normalization process was performed for the data and then risk assessment using the tool models and calculating evaluation measures (Acuracy, Precession, Recall, F1-Score, Confusion Matrix) based on estimated evaluation, real evaluation, comparison between models and highlighting the best among the models for each set of data groups.

A comparison was made between the tool models using evaluation scales for each data set to clarify the model that gives the best performance. In the dr3l data set, the best performance was for the MLP artificial neural network, as the Accuracy was 99.84%, while in the dr5l data set, the best performance was for the LGBM light gradient enhancement machine algorithm. The accuracy was 99.74%, while the data vulnerability data set had the best performance for the CatBoost algorithm, as the accuracy was 99.38%.

The performance of the tool's models was also compared with previous work that used the same metrics with the same 50,000-sample data set

that included four models (Decision tree (DT), K-Nearest Neighbors, and Support vector machine(SVM) , MLP), and the results showed the supority of the tool models over previous work models according to the evaluation scales (Accuracy, Precession, Recall, F1-Score, Confusion Matrix)..