



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

## تصنيف مخاطر البرمجيات باستخدام التعلم الآلي

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

من قبل

رباب طارق محمود الطائي

بإشراف  
أ.د. إبراهيم احمد صالح الحديدي

## المستخلص

تعد مخاطر البرمجيات من أهم التحديات التي يواجهها قطاع تكنولوجيا المعلومات وتطوير البرمجيات. فهذه المخاطر قد تؤدي إلى تعطل النظام أو تسريب البيانات الحساسة أو فقدان المعلومات، أو حتى تهديد أمن المستخدمين. لذا، فإن فهم وتحليل وتصنيف هذه المخاطر يؤدي دورًا حاسمًا في ضمان جودة البرمجيات وحماية الأنظمة من التهديدات.

تهدف هذه الرسالة إلى تصنيف مخاطر البرمجيات باستخدام خوارزميات التعلم الآلي. تم استخدام مجموعة بيانات (Software Requirement Risk Prediction Dataset) والتي تحتوي 299 متطلبًا برمجيًا يضم كل متطلب سمات متطلبات البرمجيات المتعلقة بالمتطلبات والمخاطر وفئات المشروع، فضلاً عن القيم الرقمية للاحتتمالية وحجم المخاطر وتأثيرها على عدد الوحدات ومدة التثبيت وتكلفة الإصلاح والأولية ومستوى المخاطر التي تحيط بها حيث تمت اجراء معالجة أولية للبيانات إذ تم تطبيع البيانات باستخدام خوارزمية (min-max normalization)، ومن ثم إختيار السمات باستخدام معامل الارتباط (Correlation) لتكون جاهزة لتطبيقها على خوارزميات التعلم الآلي المتمثلة بخوارزمية شجرة القرار (Decision Tree) والغابة العشوائية (Random Forest) وتعزيز التدرج (Gradient Boosting) والجار الأقرب (K- Nearest Neighbour) والانحدار اللوجستي (Logistic Regression) و الخوارزميات المحسنة (Gradient Boosting + Decision Tree + Logistic Regression) و (Gradient Boosting + Decision Tree + K- Nearest Neighbour).

تم استخدام عدة خوارزميات في هذا البحث بما في ذلك شجرة القرار (Decision Tree) ، والغابة العشوائية (Random Forest) ، وتعزيز التدرج (Gradient Boosting) ، والجار الأقرب (K-Nearest Neighbor) ، والانحدار اللوجستي (Logistic Regression) ، بالإضافة إلى خوارزميات محسنة تجمع بين هذه الخوارزميات. نتائج الاختبار أظهرت أن شجرة القرار والغابة العشوائية وتعزيز التدرج يحققون أعلى أداء في تصنيف مخاطر البرمجيات بنسبة نجاح تقدر بـ 98%، تليها الجار الأقرب بنسبة نجاح تبلغ 82%، والانحدار اللوجستي بنسبة نجاح 74%.

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



# **Software Risk Classification Using Machine Learning**

**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  
Rabab Tariq Mahmood AL-Taee**

**Supervised by  
Prof. Dr. Ibrahim A. Saleh Al- Hadidi**

## **Abstract**

The risks associated with software development are among the most critical challenges faced by the information technology and software development sector. These risks can lead to system failures, data breaches, information loss, and even threaten user security. Therefore, understanding, analyzing, and predicting these risks play a crucial role in ensuring software quality and protecting systems from threats.

This message aims to classify software risks using machine learning algorithms. A dataset called the "Software Requirement Risk Prediction Dataset" was utilized, consisting of 299 software projects. Each project included features related to software requirements, risks, project categories, as well as numerical values representing probability, risk magnitude, impact on the number of units, installation duration, repair cost, priority, and the level of associated risks. The data underwent initial preprocessing, including normalization using min-max normalization, and feature selection using correlation coefficient to prepare it for application with machine learning algorithms.

Several machine learning algorithms were employed in this research, including Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbor, Logistic Regression, and optimized combinations of these algorithms (GB+DT+LR and GB+DT+KNN). The results of the experiments demonstrated that Decision Tree, Random Forest, and Gradient Boosting achieved the highest performance in predicting software risks with an accuracy rate of approximately 98%. They were followed by K-Nearest Neighbor with an accuracy rate of 82%, and Logistic Regression with an accuracy rate of 74%.