وزارة التعليم العالي والبحث العلمي

جامعة الموصل

كلية علوم الحاسوب والرياضيات

قسم البرمجيات

كشف الحالات الشاذة في انظمة البرمجيات بناءً على التعلم التجميعي

رسالة مقدمة

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

البرمجيات

من قبل

رغده ازاد حسن حسن

بإشراف

أ.د. ابراهيم احمد صالح احمد

١٤٤٦هـ

٢٠٢٥ م

المستخلص

تُعدُّ البرمجيات عنصرًا أساسيًا في مختلف مجالات الحياة، اذ تؤدي دورًا حيويًا في تعزيز الإنتاجية وتسهيل المعالجات اليومية في القطاعات الاقتصادية والصحية والتعليمية والصناعية. ومع تزايد الاعتماد على البرمجيات، أصبح ضمان جودة البرمجيات وموثوقيتها هدفًا أساسيًا. ومع ذلك، يُعدُّ تحقيق هذا الهدف تحديًا كبيرًا بسبب القيود الزمنية والموارد المالية المحدودة. لهذا السبب، تلجأ شركات تطوير البرمجيات بشكل متزايد إلى تقنيات تعلم الآلة للتنبؤ بالحالات الشاذة، بهدف تقليل المخاطر وضمان أداء مستقر وفعّال للأنظمة البرمجية.

تهدف هذه الرسالة بتقديم نموذجا تنبئيا لاكتشاف حالات الشذوذ في الانظمة البرمجية ، ويسهم هذا النهج بشكل فعّال في تحسين دقة التنبؤ نظرًا لتعقيد المهمة وتنوع البيانات المستخدمة. تتميز النماذج القائمة على تعلم الآلة بقدرتها على دمج مجموعة متنوعة من الخوارزميات والمعالجات، مما يؤدي إلى تعزيز أدائها بشكل ملحوظ. ومن خلال التعلم المشترك بين النماذج المختلفة، يمكن تعزيز كفاءة عملية التنبؤ بشكل كبير. تُعدُّ هذه الاستراتيجية أساسية لتحقيق الاستقرار وتحسين جودة البرمجيات، مما يضمن تقديم أنظمة أكثر موثوقية وكفاءة.

تم اقتراح بناء نموذج تنبؤي فعال يقوم بدمج نتائج عدة خوارزميات تعلم آلي لتحقيق دقة عالية في اكتشاف الحالات الشاذة التي تُعدُّ من أبرز التحديات التي تواجه مطوري البرمجيات؛ اذ تؤدي إلى أعطال مفاجئة وانخفاض في كفاءة النظام. اعتمدت الدراسة على جمع بيانات متعلقة بأداء أنظمة برمجية متنوعة وتحليلها وقد استعملت ثلاثة انواع من مجاميع البيانات من مواقع مختلفة وتشمل:المجموعة الاولىالشذوذ البرمجية وتشمل (ar1,ar3,ar4,ar5,ar6) ، المجموعة الثانية هي بيانات شذوذ الاداءPerformance وتشمل (elasticsearch,structr,xchange)، وأخيراًمجموعة بيانات الشذوذ الأمنية Secuity وتشمل(ambari,camel,derby,wicket) فضلاً عن ذلك تم أجراء معالجة مسبقة للبيانات لاختيار الميزات وتحسين النموذج لتحقيق الأداء الأمثل، ثم تطبيق خوارزميات التعلم التجميعي مثل التعبئة(Bagging) والتكديس(Stacking) والتصويت(Voting) والتعزيز التكيفي(AdaBoosting) وتعزيز التدرج الشديد(XGBoosting)) للتنبؤ بالانماط الشاذة في هذه البيانات.

تم تقييم التنبؤ بالحالات الشاذة وحساب مقاييس لتقييم (Accuracy, Precision, Recall, F1-Score) واجراء مقارنة بين خوارزميات النموذج المقترح ففي مجموعة البيانات الخاصة بالشذوذ البرمجية ، تشير النتائج الى أن خوارزميات التعلم التجميعي تشير الى ستراتيجية فعالة في كشف الشذوذ في الانظمة

البرمجية ، اذ حققت خوارزمية التعزيز التكيفي (AdaBoosting) افضل اداء لمعظم المجاميع إذ حققت اعلى دقة 100% عند المجموعةar5، أما مجموعة بيانات شذوذ الاداء فقد حققت خوارزمية التعبئة (Bagging) افضل اداء لمعظم المجاميع إذ بلغت دقة التنبؤ 99%عند المجموعة structr . وحققت خوارزمية التصويت (Voting) افضل اداء لمعظم المجاميع الخاصة بالشذوذ الامنية . وأجريت مقارنة النموذج المقترح مع نماذج اخرى لدراسات سابقة استخدمت مجاميع البيانات نفسها وتبين تفوق النموذج المقترح عليها اعتماداً على مقاييس التقييم المستخدمة .

Ministry of Higher Education and Scientific Research

University of Mosul

College of Computer Science and Mathematics

Software Department

Detection for Anomaly States in Software Systems based on Ensemble 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

Raghda Azad Hasan Hassan

Supervised by

Prof. Dr. Ibrahim Ahmed Saleh Ahmed

2025 A.D. 1446 A.H.

Abstract

Software is an essential element in various areas of life, as it plays a vital role in enhancing productivity and facilitating daily transactions in the economic, health, educational and industrial sectors. With the increasing reliance on software, ensuring software quality and reliability has become a primary goal. However, achieving this goal is a major challenge due to time constraints and limited financial resources. For this reason, software development companies are increasingly resorting to machine learning techniques to predict anomalies, with the aim of reducing risks and ensuring stable and efficient performance of software systems.

This thesis aims to present a predictive model for detecting anomalies in software systems, and this approach effectively contributes to improving the accuracy of prediction due to the complexity of the task and the diversity of data used. Machine learning-based models are characterized by their ability to integrate a variety of algorithms and processors, which leads to a significant enhancement in their performance. Through joint learning between different models, the efficiency of the prediction process can be greatly enhanced. This strategy is essential for achieving stability and improving software quality, ensuring the provision of more reliable and efficient systems.

An effective predictive model was proposed that integrates the results of several machine learning algorithms to achieve high accuracy in detecting anomalies, which are among the most prominent challenges facing software developers, as they lead to sudden failures and a decrease in system efficiency. The study relied on collecting and analyzing data related to the performance of various software systems. Three types of data sets were used from different locations, including: the first set of software anomalies, which includes (ar1, ar3, ar4, ar5, ar6), the second set is performance anomaly data, which includes (elasticsearch, structr, xchange), and finally the security anomaly data set, which includes (ambari, camel, derby, wicket). In addition, the data was preprocessed to select features and improve the model to achieve optimal performance, then ensemble learning algorithms such as bagging, stacking, voting, AdaBoosting, and XGBoosting were applied to predict anomaly patterns in this data.

The prediction of anomalies was evaluated and the evaluation metrics (Accuracy, Precision, Recall, F1-Score) were calculated and a comparison was made between

the algorithms of the proposed model. In the dataset of software anomalies, the results indicate that the ensemble learning algorithms indicate an effective strategy in detecting anomalies in software systems, as the adaptive boosting algorithm (AdaBoosting) achieved the best performance for most groups, as it achieved the highest accuracy of 100% for the ar5 group. As for the performance anomaly dataset, the bagging algorithm achieved the best performance for most groups, as the prediction accuracy reached 99% for the structr group. The voting algorithm achieved the best performance for most groups of security anomalies. The proposed model was compared with other models of previous studies that used the same datasets, and it was shown that the proposed model outperformed them based on the evaluation metrics used.