



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

التنبؤ بالعيوب البرمجية 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

طارق نجم عبيد عيدان

بإشراف

أ.د. صفوان عمر حسون

المستخلص

تلعب البرمجيات دوراً حيوياً في جميع جوانب حياتنا اليومية، وتحديداً في مجالي الطب والصناعة. ومن أجل تصميم برمجيات عالية الجودة والموثوقية، وتجنب المخاطر المادية والبشرية الناتجة عن الأخطاء البرمجية، يعتبر الأمر تحدياً كبيراً نظراً للقيود الزمنية والميزانية المحدودة. لهذا السبب، لجأت العديد من شركات تطوير البرمجيات إلى استخدام التعلم الآلي لتنبؤ واكتشاف العيوب البرمجية، باستخدام المعرفة المستفادة من سجلات المشاريع السابقة والبيانات المتاحة. هذا النهج يسمح لهم باتخاذ قرارات أكثر دقة وفعالية خلال مراحل تطوير البرمجيات، وبالتالي يساهم في تحسين جودة وأمان البرمجيات المنتجة.

تم اقتراح نموذج يساعد في زيادة دقة التنبؤ بسبب تعقيد المهمة وتباين البيانات، ويمكن النماذج استغلال مجموعة متنوعة من الخوارزميات والمعالجات، مما يؤدي إلى تحسين أدائها بشكل كبير. بالإضافة إلى ذلك، يمكن جمع البيانات من مصادر متعددة، مما يقلل من تأثير تقلبات البيانات ويزيد من استقرار النتائج. تمتلك هذه النماذج مرونة عالية تسمح بتحسينها وتعديلها بسهولة بناءً على الحاجة والبيانات الجديدة، وبالتعلم المشترك بين هذه النماذج، يمكن تعزيز قدرتها على التنبؤ بالعيوب بشكل أفضل وبدقة أكبر. هذه الاستراتيجية تعد أساسية لتحسين جودة البرمجيات وزيادة كفاءة التنبؤ بالعيوب، مما يساهم في تقديم برمجيات أكثر جودة وموثوقية.

تم استخدام مجموعة بيانات مستودع MDP التابع لوكالة ناسا والمتخصصة بالتنبؤ بالعيوب البرمجية في هذا السياق وتم إجراء اختيار السمات ذات الأهمية وموازنة مجموعة البيانات من ضمن المعالجة الأولية وتطبيق خوارزميات التعلم الآلي التجميعي (التعبئة Bagging)، وتعزيز التدرج Gradient Boosting، وتعزيز التدرج الفائق Extreme Gradient Boosting، والتعزيز الفئوي categorical Boosting، والمكدس Stacking) ومن ثم تقييم التنبؤ بالعيوب البرمجية وحساب مقاييس التقييم (Accuracy، Precision، Recall، F1-score) وإجراء مقارنة بين خوارزميات النموذج المقترح ففي مجموعة بيانات PC1 فحققت افضل اداء لخوارزمية التعزيز التدرج بدقة تقدر 0.97759 ، بينما لمجموعة بيانات (PC2,PC3,PC4,PC5) كان افضل أداء لخوارزمية المكدس بدقة تقدر ((0.9872),PC5(0.97297),PC4(0.94819),PC3(0.98675),PC2))، وتم إجراء مقارنة لنفس مجموعة البيانات مع مجموعة من الدراسات السابقة فظهر تفوق النموذج المقترح حسب مقاييس التقييم المستخدمة.

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



Software defect prediction 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
Tariq Najm Obad Edan**

**Supervised by
Prof. Dr. Safwan Omar Hasoon**

2023 A.D.

1445 A.H.

Abstract

Software plays a vital role in all aspects of our daily lives, specifically in the fields of medicine and industry. In order to design high-quality and reliable software, and avoid physical and human risks resulting from software errors, it is considered a major challenge due to time constraints and limited budget. For this reason, many software development companies have turned to using machine learning to predict and detect software defects, using knowledge learned from previous project logs and available data. This approach allows them to make more accurate and effective decisions during the software development phases, and thus contributes to improving the quality and security of the software produced.

A model has been proposed that helps increase prediction accuracy due to task complexity and data variability, and the models can exploit a variety of algorithms and processors, significantly improving their performance. In addition, data can be collected from multiple sources, which reduces the impact of data fluctuations and increases the stability of the results. These models have high flexibility that allows them to be easily improved and modified based on need and new data, and by joint learning between these models, their ability to predict defects better and with greater accuracy can be enhanced. This strategy is essential to improve software quality and increase the efficiency of defect prediction, which contributes to delivering more quality and reliable software.

The NASA MDP repository dataset, which is specialized in predicting software defects, was used in this context. Selection of important features and balancing of the dataset were carried out through pre-processing and application of ensemble machine learning algorithms (Bagging, Gradient Boosting, and Extreme Gradient Boosting), categorical boosting, and stacking) and then evaluate the prediction of software defects and calculate evaluation metrics (Accuracy, Precision, Recall, F1-score) and conduct a comparison between the algorithms of the proposed model. In the PC1 data set, it achieved the best performance of the gradient boosting algorithm with an accuracy of 0.97759, while For a data set (PC2, PC3, PC4, PC5), the best performance of the stack algorithm was estimated at an accuracy of (PC2(0.98675), PC3(0.94819), PC4(0.97297), PC5(0.9872)), and a comparison was made for the same data set with a group of studies. The superiority of the proposed model was demonstrated according to the evaluation criteria used.