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

التنبؤ بقابلية صيانة البرمجيات بناءً على التعلم الآلي

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

من قبل
مازن محمد إسماعيل الطائي

بإشراف

أ.د. ابراهيم احمد صالح الحديدي

1445 هـ 2023 م

المستخلص

تُعدُّ الصيانة مرحلة متكررة في دورة تطوير البرمجيات تتضمن إضافة الميزات الجديدة، وحذف الميزات غير المرغوب فيها، وتصحيح الخطأ، والتكيف مع الحالة الجديدة، غالباً ما تكون مرحلة الصيانة من المراحل الأكثر تكلفة واستهلاكاً للوقت. تهدف جودة البرمجيات إلى تطوير منهجيات لإنتاج برمجيات عالية الجودة وبتكاليف أقل والتأكد من أن المنتج يلبي توقعات المستخدم ومتطلباته، إذ تُعدُّ قابلية صيانة البرمجيات واحدة من خصائص الجودة التي تحدد المسار الذي يمكن أن تكون عليه هذه التعديلات لإعطاء نتائج أفضل.

في هذه الرسالة تم توليد مجموعة البيانات الخاصة بالتنبؤ بقابلية الصيانة البرمجيات، وتم ايضاً استخدام مجموعة البيانات ثم بناء الأداة للتنبؤ بقابلية الصيانة باستخدام خوارزميات (شجرة القرار-Decision Tree) ((DT)، الغابة العشوائية ((Random Forest-(RF)) ، الانحدار اللوجستي (Logistic Regression – (LR))، الجار الأقرب ((K-Nearest Neighbor-(KNN)) ، آلة متجه الداعمة (Support Vector Machine – (SVM)) ، بايز البسيط ((Naïve Bayes-(NB)) ، تم إجراء سلسلة من المعالجات المسبقة لمجموعتي البيانات المستخدمة والمولدة والتي تشمل تنظيف البيانات، وتطبيع البيانات باستخدام تقنية (Min-Max Normalization) ، وموازنة البيانات باستخدام الزيادة العشوائية (Random Oversampling) وبنسبة 60% لفئة الاقلية لمجموعة البيانات المستخدمة ونسبة 30% لفئة الاقلية لمجموعة البيانات المولدة، واستخلاص الميزات باستخدام خوارزمية صقور هاريس (Harris Hawks Optimization Algorithm) وتطبيقها على مجموعة البيانات المستخدمة فقط.

تم ضبط المعاملات باستخدام طريقة البحث الشبكة (Grid Search) لتحسين أداء خوارزميات التعلم الآلي، أثبت هذا النهج فعاليته في تحديد أفضل المعاملات لكل خوارزمية. بعد ذلك ، تم تطبيق الخوارزميات واستخدام مجموعة من مقاييس الأداء (الدقة (Accuracy)، معدل الاستدعاء (Recall)، الضبط (Precision)، درجة F1 (F1-Score)، ومصفوفة الارتباك (Confusion Metrics)، منحنى خاصية تشغيل الاستقبال (Receiver operating characteristic curve – (ROC))، لتقييم أداء الخوارزميات، ففي مجموعة البيانات المستخدمة قد تبين أن خوارزمية الغابة العشوائية هي أفضل

الخوارزميات المستخدمة في عملية التنبؤ فقد حصلت على دقة بنسبة 99%، بينما حققت كل من شجرة القرار وخوارزميات الجار الأقرب دقة تصل إلى 96٪ وكانت أقل بقليل من خوارزمية الغابة العشوائية، وحققت خوارزمية الانحدار اللوجستي دقة بلغت 84٪، بينما حققت خوارزمية آلة متجه الدعم دقة 93٪. وكانت خوارزمية بايز البسيط أقل دقة وبلغت 84٪، وفي مجموعة البيانات المولدة ايضاً كانت خوارزمية الغابة العشوائية هي أفضل الخوارزميات المستخدمة في عملية التنبؤ وبنسبة دقة 99% ، بينما حققت خوارزمية شجرة القرار دقة بلغت 98% وحققت خوارزمية الجار الأقرب دقة بلغت 98% وكانت قريبة من دقة الغابة العشوائية، وحققت خوارزمية الانحدار اللوجستي دقة بلغت 70٪، وحققت خوارزمية آلة متجه الدعم دقة 95٪، وكانت خوارزمية بايز البسيط أقل دقة وبلغت 66٪.

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

Software Maintainability Prediction Based on Machine Learning

**A Thesis Submitted to the Council of the College of
Computer Sciences and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master of Sciences
in
Software**

By

Mazin Mohammed Ismael Al.Taiee

Supervised by

Prof.Dr. Ibrahim Ahmed Saleh Al.Hadedi

2023 A.D. **1445 A.H.**

Abstract

Maintenance is an iterative stage in the software development life cycle that includes adding new features, deleting unwanted features, correcting the error, and adapting to the new state. The maintenance stage is often one of the most expensive and time-consuming stages. Software quality aims to develop methodologies for producing high-quality software at lower costs and to ensure that the product meets the user's expectations and requirements. Software maintainability is one of the quality characteristics that determines the course on which these modifications can be to give better results.

In this thesis, the data set for predicting the maintainability was generated, and the data set was also used and then the tool was built to predict the maintainability using the algorithms (Decision Tree-(DT)), Random Forest-(RF)), regression Logistic Regression-(LR)), K-Nearest Neighbor-(KNN)), Support Vector Machine-(SVM)), Naïve Bayes-(NB)), were performed. A series of pre-treatments for the used and generated data sets, which include data cleaning, data normalization using the (Min-Max Normalization) technique, balancing the data using random oversampling, 60% for the minority class of the used data set, and 30% for the minority class for the generated data set, Extract features using Harris Hawks Optimization Algorithm and apply them to the used dataset only.

The parameters were tuned using the Grid Search method to improve the performance of the machine learning algorithms. This approach has proven effective in identifying the best parameters for each algorithm. After that, the algorithms were applied and a set of performance metrics were used (Accuracy, Recall rate, Precision, F1-Score, Confusion Metrics), Receiver operating characteristic curve characteristic curve – (ROC)), to evaluate the performance of the algorithms. In the data set used, it was found that the random forest algorithm is the best algorithm used in the prediction process, as it obtained an accuracy of 99%, while both the decision tree and the nearest neighbor algorithms achieved an accuracy of up to 96% which was slightly lower than the random forest algorithm, the logistic regression algorithm achieved an accuracy of 84%, while the support

vector machine algorithm achieved an accuracy of 93%. The simple Bayes algorithm was less accurate and amounted to 84%, and in the generated data set also the random forest algorithm was the best algorithm used in the prediction process with an accuracy of 99%, while the decision tree algorithm achieved an accuracy of 98% and the nearest neighbor algorithm achieved an accuracy of 98% which It was close to the accuracy of the random forest, and the logistic regression algorithm achieved an accuracy of 70%, while the support vector machine algorithm achieved an accuracy of 95%, and the simple Bayes algorithm was the least accurate and reached 66%.