Software Bug Prediction Using Supervised Machine Learning Algorithms

S. Delphine Immaculate
Department of Computer Science & Engineering
Mookambigai College of Engineering
Tiruchirappalli, India
delphine.rajesh@yahoo.co.in

M. Farida Begam
Department of Information Science and Engineering
CMR Institute of Technology
Bengaluru, India
farida.b@cmrit.ac.in

M. Floramary
Department of Computer Science & Engineering
Mookambigai College of Engineering
Tiruchirappalli, India
floramary363@gmail.com

***Abstract*—Machine Learning algorithms sprawl their application in various fields relentlessly. Software Engineering is not exempted from that. Software bug prediction at the initial stages of software development improves the important aspects such as software quality, reliability, and efficiency and minimizes the development cost. In majority of software projects which are becoming increasingly large and complex programs, bugs are serious challenge for system consistency and efficiency. In our approach, three supervised machine learning algorithms are considered to build the model and predict the occurrence of the software bugs based on historical data by deploying the classifiers Logistic regression, Naïve Bayes, and Decision Tree. Historical data has been used to predict the future software faults by deploying the classifier algorithms and make the models a better choice for predictions using random forest ensemble classifiers and validating the models with K-Fold cross validation technique which results in the model effectively working for all the scenarios.**

Keywords —bug; classifier; cross validation; defect; machine learning; software metrics.

I. INTRODUCTION

A. Machine Learning

Machine learning provides ability predict the occurrence of an event based on historical data without being explicitly programmed. Machine learning focuses on the output variable and try to look for patterns within the data and predicts the final outcome. There are two kinds of machine learning algorithms. Supervised, Unsupervised machine learning algorithms. In supervised algorithms, we will have an output variable decided upfront and we try to look for pattern similarities between the dependent and independent variables. In un supervised machine learning algorithms, one can pass the data and predict the final outcome variable. In case of un supervised machine learning algorithms, the concept of dependent variable is not there. The algorithms uses the complete data and creates the final outputs. Machine learning algorithms help to track real time behaviors that can be used as inputs for companies to reduce operational cost and gain better result. Fault prone development of software components will definitely leads to loosing of business. With the help of machine learning algorithms, uncertainty can be predicted with certain confidence up front which leads to trimming of chaos or disorganization and increases the momentum of automation in SDLC.

B. Defect Prediction Using Machine Learning Algorithms

Predicting the bug occurrence and understanding the process flow of SDLC are the key factors which leads to success criteria. Defect prediction using Machine Learning algorithm can be applied for any stage of SDLC such as , identification of current problems, plan, design, build, test, deploy and maintain and also any model type of SDLC such as Waterfall Model, Agile Model, Iterative Model, V-Shaped Model, Big Bang Model and Spiral Model. Machine learning algorithms and statistical analysis actually helps to guess that a code is buggy. Improved and better approaches in the development will increase the quality of the software.

C. Defect Life Cycle

Predicting the occurrence of bug and understanding the defect life cycle are very essential and mandatory which signifies the importance of predicting the bug earlier in the SDLC. It is helpful to avoid the time, effort and cost spent in detecting and fixing the defects. One of the key issues the modern software industry is facing is the number of bugs raised during the development cycle. This leads to an addition time to final delivery of the product and total increase in the operation cost . It is estimated to be 10% average operation cost increase across the projects because of the bugs that happens during code development and production of the software. As this is the hour of the need, we try to address the defect predictions using advanced machine learning algorithms. The model uses parameters from multiple dimensions and try to predict the occurrence of the bug. The model can be used as input for the occurrence of the bug before a tester identifies. Hence, the developer can retrospect the code and fix the issues with in short duration of the time. The model is well suited for most famous software development frameworks like waterfall, agile, V-shape, spiral and etc. By keeping the different kinds of software development modeling frameworks, we have constructed this final machine learning models. This acts as a universal model for all kind of defect prediction algorithms. Understanding the defect management life cycle is one of the key aspects. Defect life cycle or Bug Life cycle shown in Fig. 1 gives the insight of the states of the defect it pass through in its valid tenure.

978-1-5386-9319-3/19/$31.00 ©2019 IEEE