Chapter 3

Problem Formulation
3.1 Introduction

The success of any software system is entirely depend on the accuracy of the results of the system and whether it is without any flaws, software defect prediction problems have an extremely beneficial research potential. Caper, [56] investigated the software bugs or defects have rendered major contributions towards technical explanations for software project failure. Mining of software repositories have several research challenges to be addressed. New software bug prediction models need to be designed, effective software defect metrics need to be synthesized and provided them as inputs to various data mining techniques for extracting classified information in order to envisage the software faults in new software versions and also more developed methods are needed to reduce software cost overruns[57].

The software metrics (for example, process metrics and product metrics) lie at the core of bug prediction models. The main objective of any organization is to have defect free software. In fact, earlier detection of defect would save time and cost of the system. The above statement clearly describes that we need to identify the finest machine learning model for software defect prediction for which various performance parameters are available, such as, accuracy, mean square error, and correlation and R-Squared to compare with other diverse models. Data is an important part of the system. One of the biggest challenges is to acquire the right dataset and where in to categorize the dependent and independent variables. The more the data, the more complex will the system become and more probability of the defect appearing. Hence, it is always safer to remove the insignificant variables from the dataset and reduce the independent variable using the feature selection technique. The insignificant variables have a negligible impact to detect the software bug. There are various types of feature selection techniques that are available to derive the significant and insignificant variables from the dataset. For the research, the Wrapper and Filter method used in Feature Selection technique are taken to find the imperative variable from the publicly available Promise Repository.

A machine-learning technique can be classified as a supervised learning and unsupervised learning. Unsupervised learning works on a hidden data and consists of clustering; association analysis, hidden Markov model. The supervised learning is used when there is a requirement to train the model for prediction. It comprises of regression and classification. Most of the researches are based on classification techniques and only a few have been done using regression
techniques used in a machine learning model. Research has been carried out using the regression technique and machine learning model used are Linear Regression, Decision Tree, Random Forest, Neural Network, Support Vector Machine and Decision Stump. Performance parameters were applied on these machine learning models to attain the finest model to predict the software defect.

Machine learning techniques, Naïve Bayes, Decision Stump, SVM-Polykernel, and SVM-RBF were used by Selvaraj et.al. [58] to calculate the software defect using classification method and concluded that SVM-Polykernel attain the finest performance and the dataset taken was Promise Repository. Yogesh et.al [59] has used the NASA dataset for the experiment and found that the SVM model offers the best accuracy out of logistic regression and the decision tree to predict the software defect. Elish et.al [60] too used the NASA dataset and applied eight statistical and machine learning models to predict the defect prone module and have declared that the SVM as the best possible model as compared with Logistic Regression (LR), K-Nearest Neighbor (KNN), Multi-Layer Perceptrons (MLP), Radial Basis Function (RBF), Bayesian Belief Networks (BBN), Naive Bayes (NB), Random Forests (RF) and Decision Trees (DT).

Most of the Researchers have concluded that the Support Vector Machine (SVM) is by far the best model to predict the software defect yet some of them have employed classification technique and applied on various datasets. The research gap was to apply the machine learning model using regression technique and to check whether in other datasets, it also achieving the same results.