SOFTWARE COST PREDICTION USING ML TECHNIQUES

MADHAV SHARMA

madhav.209302295@muj.manipal.edu

***Abstract*—Software cost estimation is a crucial process for the success of software projects. Accurate estimation results can help project managers and software engineers to plan and manage resources, understand project progress and phases, and manage human resources, assets, software, data, and feasibility study. Developing an accurate cost estimation model for a software project, however, is challenging. Machine learning can play a vital role in achieving these objectives.**

This research paper aims to apply three different machine learning models (SVR, Random Forest, Decision Tree Regressor) on two different datasets (COCOMO81 and Desharnais) to predict software cost. The performance of each model and dataset is evaluated using MMRE and MdMRE. The results are used to determine the best machine learning model for predicting accurate software cost and to discuss its strengths and limitations. Potential ways to improve the accuracy of software cost prediction using machine learning are also explored. This research highlights the importance of accurate software cost estimation and the potential of machine learning to assist project managers and software engineers in achieving this goal.

I. INTRODUCTION

The accurate estimation of software project costs, resources, and effort is crucial for the success of software development projects. However, despite the availability of various estimation techniques and models, inaccurate estimations and budget overruns remain a persistent problem for software professionals, clients, and stakeholders. One of the well-established estimation models, COCOMO, has some limitations such as dependence on historical project data, inability to estimate in all software development life cycle phases, and requirement of a large amount of input data. Hence, this research aims to explore the use of machine learning algorithms for more accurate software project effort estimation using COCOMO model datasets. The research will compare the performance of machine learning models with the COCOMO model and discuss the strengths and limitations of using machine learning algorithms for software project estimation.

(

II. LITERATURE REVIEW

A review of existing literature on software cost prediction reveals a wide range of approaches and techniques that have been used in previous research. These include statistical methods, rule-based techniques, and machine learning algorithms such as Random Forests,SVM, decision trees, and others. Previous studies have shown that machine learning of software fault prediction. However, there are still research techniques that have the potential to significantly improve the accuracy and efficiency of software cost prediction. However, there are still research gaps in terms of identifying the most effective machine learning algorithms, feature selection techniques, and evaluation metrics for software cost prediction aims to address these gaps and contribute to the existing body of knowledge in this field

III. METHODOLOGY

To develop a software cost prediction model using machine learning, we first gather a comprehensive dataset of software metrics from various projects. These metrics may include code complexity, size. Programmer capability, experience and other relevant features. We preprocess the data to handle missing values, normalize the features, and remove any irrelevant or redundant features. After this, we apply feature selection techniques such as correlation analysis, information gain, or recursive feature elimination to identify the most relevant features that impact fault prediction. Next, we apply various machine learning algorithms, such as Random Forest,SVM, decision trees, and others, to build predictive models. We divide the dataset into training and testing sets and use cross-validation techniques such as k-fold cross-validation to assess the models' performance. To evaluate our approach's effectiveness, we measure metrics such as MMRE, MdMRE, and PRED(0.25).

TABLE I Number of attributes and projects in different datasets

Dataset	Attributes	Number of Projects
COCOMO NASA1	17	60
COCOMO 81	16	63
DESHARNAIS	12	81

TABLE II. COMPARISION OF VARIOUS ML MODEL WITH COCOMO 81 DATASET

Algorithm	MMRE	MdMRE	PRED(0.25)
SVM	1.2060	0.685	5.263
Random Forest	0.764	0.815	5.263
Decision Tree	0.75	0.839	5.263

TABLE III. COMPARISION OF VARIOUS MACHINE LEARNING MODELS WITH DESCHARNAIS DATASET

Algorithm	MMRE	MdMRE	PRED(0.25)
Random Forest (Grid Search CV)	0.882	0.366	40.0
RF(Random Search)	0.89782254	0.30826	40.0
Decision Tree	0.8774700	0.504942	36.0

TABLE IV. Evaluation metrics and best hyper parameters resulted from grid search method

Datasets	MMRE	MdMRE	PRED(0.25)	Best Parameters	
				Max_features	n_estimators
COCOMO 81	0.7643	0.81	5.26	10	1200
Desharnais	0.89	0.36	40.0	1	400

IV. RESULTS

The results of our research show that our proposed approach using machine learning techniques for software cost prediction yields promising results. The accuracy, precision, recall, and MMRE measure of our predictive models are significantly improved compared to traditional methods. The results of the experiments found clear support that Decision Trees and Random Forest algorithms impressively give consistent results with the COCOMO datasets regardless on the number of effort attribute used. Furthermore, we compare our results with existing research in the field of software cost prediction, demonstrating the effectiveness and potential of our approach. Our findings highlight the advantages of utilizing machine learning techniques in software cost prediction and their potential for improving the quality and reliability of software applications.

V. CONCLUSIONS

In this study, we aimed to investigate the application of Random Forest for Software Development Effort Estimation. To this end, we used three widely-used SDEE datasets: Albrecht, Desharnais, and COCOMO81. The experiments were conducted to determine the impact of hyperparameters on the estimation model. In addition, we conducted exhaustive grid search and randomized search to identify the optimal hyperparameters for the Random Forest estimation model. We then compared the performance of the Random Forest model MdMRE, and PRED(25). Our findings revealed that tuning the hyperparameters resulted in improved accuracy of the RF model, leading to the best performing RF model.and the Regression Tree model using a 30% holdout validation method, based on three evaluation metrics: MMRE,

ACKNOWLEDGMENTS

The author extends heartfelt thanks to all the researchers involved in this study. The author would like to express sincere appreciation to Dr. Sudhir Sharma for his invaluable guidance and support throughout the research. The author is also grateful to Dr. Sharma for his timely and constructive feedback on the manuscript drafts, which helped to improve the overall quality of this project. The author considers working under Dr. Sharma's mentorship to have been an excellent learning experience.

REFERENCES

[1] Zakrani, Abdelali & Mustapha, Hain & Namir, Abdelwahed. (2018). Software Development Effort Estimation Using Random Forests: An Empirical Study and Evaluation. International Journal of Intelligent Engineering and Systems. 11. 300-311. 10.22266/ijies2018.1231.30. I. Boglaev, “A numerical method for solving nonlinear integro-differential equations of Fredholm type,” *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.

[2] Z. Noor Azura et al., "Software Project Estimation with Machine Learning," in International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.