SOFTWARE COST PREDICTION USING ML TECHNIQUES

Madhav Sharma*1*, Dr Sudhir Sharma *2*

1 Manipal University Jaipur 1, Jaipur-Ajmer Express Highway, Dehmi Kalan, Jaipur, India

2 Manipal University Jaipur 1, Jaipur-Ajmer Express Highway, Dehmi Kalan, Jaipur, India

Abstract

This research investigates the effectiveness of Random Forest, a machine learning technique, for software development effort estimation. By utilizing two widely used datasets, Desharnais and COCOMO81, the models are trained and evaluated. The study focuses on optimizing the Random Forest model's hyperparameters through exhaustive grid search and randomized search. A 30% holdout validation method is employed to compare the performance of the optimized Random Forest model against the Regression Tree model, using three evaluation metrics: Mean Magnitude of Relative Error (MMRE), Median Magnitude of Relative Error (MdMRE), and Prediction within 25% of the actual effort (PRED(25)).The findings demonstrate that tuning the hyperparameters significantly enhances the accuracy of the Random Forest model, surpassing the performance of the Regression Tree model. This research underscores the potential of machine learning techniques, particularly Random Forest, in accurately estimating software development effort. The improved accuracy, precision, and recall compared to traditional methods offer valuable insights for project managers and software engineers, aiding them in effective resource planning and allocation.

Keywords [[1]](#footnote-1)

Machine Learning; Random Forest; Regression Tree; Software Development Effort Estimation; Hyperparameter Optimization; MMRE; MdMRE; PRED (25).

# 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. Despite the availability of various estimation techniques, inaccuracies and budget overruns remain persistent issues for software professionals and stakeholders. COCOMO, a well-established estimation model, has limitations, including reliance on historical data, inability to estimate all development phases, and the need for a large amount of input data. To address these challenges, the software community seeks innovative, data-driven approaches and machine learning techniques that offer more adaptable and accurate solutions to software cost estimation, potentially revolutionizing estimation practices and empowering stakeholders with reliable forecasts throughout the development life cycle. 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.

# 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, and decision trees. 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.

# 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 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 various machine learning algorithms, such as Random Forest (RF), 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).

Definitions of metrics:

Table 1

Number of Attributes and projects in Different datasets

|  |  |  |
| --- | --- | --- |
| Algorithm | Attributes | Number of Projects |
| **COCOMO 81** | 16 | 63 |
| **DESHARNAIS** | 12 | 81 |

# 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 of the number of effort attribute used. Furthermore, we compared 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. We also present the best hyperparameters for Random Forest Model that resulted from grid search CV.

Table 2

Comparison of various ML models on COCOMO81 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 3

Comparison of various ML models on Desharnais Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | MMRE | MdMRE | PRED (0.25) |
| **RF (Grid Search CV)** | 0.882 | 0.366 | 40.0 |
| **RF(Random Search)** | 0.8978 | 0.3082 | 40.0 |
| **Decision Tree** | 0.8774 | 0.5049 | 36.0 |

Table 4

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | 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 | |

# Conclusions

In this study, we aimed to investigate the application of different ML Models for Software Development Effort Estimation. To this end, we used two popular SDEE datasets: Desharnais, and COCOMO81. The experiments were conducted to determine the best ML model for software cost estimation. 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 MMRE, and PRED (25). Our findings revealed that tuning the hyperparameters resulted in improved accuracy of the RF model, leading to the best performing Random Forest model. The performance of Random Forest, Decision Tree and Support Vector Regression Models was evaluated based on MMRE (Mean Magnitude of Relative Error). Random Forest and Decision Trees had the lowest MMRE values and hence were the most accurate.

A picture containing text, screenshot, font, diagram

Description automatically generated

A picture containing text, screenshot, number, diagram

Description automatically generated

# Acknowledgements

The author is sincerely grateful to all the researchers who contributed to this study, as their dedication and insights enriched the findings. Special thanks go to Dr. Sudhir Sharma for invaluable guidance throughout the research. Appreciation is also extended to friends, and family for their unwavering support.

# References

1. Zakrani, A., Mustapha, H., & Namir, A. (2018). Software Development Effort Estimation Using Random Forests: An Empirical Study and Evaluation. International Journal of Intelligent Engineering and Systems, 11, 300-311. doi:10.22266/ijies2018.1231.30.
2. Z. Noor Azura et al., "Software Project Estimation with Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.
3. Mahmood Asheeri and Mustafa Hammad, "Machine Learning Models for Software Cost Estimation," pp. 1-6, 2019. doi:10.1109/3ICT.2019.8910327.
4. Farrukh Arslan, "A Review of Machine Learning Models for Software Cost Estimation," Review of Computer Engineering Research, Conscientia Beam, vol. 6, no. 2, pp. 64-75, 2019.
5. O. Alhazmi and M. Khan, "Software Effort Prediction Using Ensemble Learning Methods," Journal of Software Engineering and Applications, vol. 13, pp. 143-160, 2020. doi:10.4236/jsea.2020.137010.
6. S. M. Satapathy, B. P. Acharya, and S. K. Rath, "Early stage software effort estimation using random forest technique based on use case points," IET Software, vol. 10, no. 1, pp. 10-17, 2016.
7. F. Arslan, "A Review of Machine Learning Models for Software Cost Estimation," Review of Computer Engineering Research, Conscientia Beam, vol. 6, no. 2, pp. 64-75, 2019.
8. S. S. Ali, J. Ren, K. Zhang, J. Wu, and C. Liu, "Heterogeneous Ensemble Model to Optimize Software Effort Estimation Accuracy," IEEE Access, vol. 11, pp. 27759-27792, 2023. doi: 10.1109/ACCESS.2023.3256533.

1. International Conference on Applied Artificial Intelligence, October 05–06, 2023, Oxford, United Kingdom

   EMAIL: [sharmamadhav208@gmail.com(Madhav](mailto:sharmamadhav208@gmail.com(Madhav) Sharma); [sudhir.sharma@jaipur.manipal.edu(Sudhir](mailto:sudhir.sharma@jaipur.manipal.edu(Sudhir) Sharma)

   ORCID: XXXX-XXXX-XXXX-XXXX (A. 1); XXXX-XXXX-XXXX-XXXX (A. 2); XXXX-XXXX-XXXX-XXXX (A. 3)

   |  |  |
   | --- | --- |
   |  | ©️  2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). |
   |  | CEUR Workshop Proceedings (CEUR-WS.org) |

   [↑](#footnote-ref-1)