Quest Journals
Journal of Software Engineering and Simulation
Volume 9 ~ Issue 4 (2023) pp: 01-07
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Analysis of Machine Learning Techniques Used In Software Cost Estimation- A Review

Beatrice O. Akumba.,[1] Nachamada Blamah[2], Iorshase Agaji[3] and Emmanuel Ogalla[3]

[1]*Department of Mathematics/Computer Science, Benue State University Makurdi, Nigeria*
[3]*Department of Computer Science, University of Jos, Nigeria*
[2]*Department of Mathematics/Statistics/Computer Science, Joseph Sarwuan Tarka University Makurdi, Nigeria*

ABSTRACT
Software project cost estimation (SCE) is the process of predicting the total cost needed to develop a software system. The prediction of software project costs and efforts are complicated most times due to unclear and uncertain requirements needed from the customers by the developers. An extensive review of some Authors work was done in SCE and the findings and analysis are summarized. It was seen that most researchers in SCE used machine learning techniques to develop their models that provided accurate predictions and estimations. The datasets used mostly for the model predictions is the COCOMO and NASA datasets respectively. The evaluation metrics used by most researchers to evaluate the machine learning SCE models developed are the Mean Magnitude of Relative Error (MMRE), Percentage Relative Error Deviation (PRED) and Root Mean Squared Error (RMSE) respectively. Meanwhile, there were a lot of limitations and recommendations registered by most Authors reviewed for the accurate software project cost/effort prediction but the most common one is the recommendation of a hybrid machine learning model to improve the accuracy problem inherent in most SCE models.

KEYWORDS: *Software Cost Estimation (SCE), Machine Learning Techniques (MLT), Software Project Management (SPM), Software Engineering, Model Performance Metrics*

***R**eceived 25 Mar., 2023; **R**evised 05 Apr., 2023; **A**ccepted 07 Apr., 2023 © The author(s) 2023.*
P**ublished with open access at **www.questjournals.org

I. INTRODUCTION

Software engineering entails the standard procedures that sheath the diverse aspects of the development of software project beginning from its establishment to its post-release maintenance. The standards enable well timed and efficient project completion despite preserving the quality at the highest level. Software project management (SPM) is the main course of action for handling software projects. The activities required for SPM process are: planning, supervising and integrating which leads to establishments of the software project. In software project management system, the estimation of the software project cost and effort are the most crucial activities as there often tends to be no background information or any previous experience about most software projects. Software project cost estimation is the process of predicting the total cost needed to develop a software system [1]. The prediction of software project costs and efforts are complicated most times due to unclear and uncertain requirements needed from the customers by the developers.

Accurate software cost estimates are important to both developers and customers. It can be used for generating request for proposals, contract negotiations, scheduling, monitoring and control. Underestimation of the costs may result in management approving proposed systems that then exceed their budgets, with underdeveloped functions and poor quality, and failure to complete on time. However, overestimating on the other hand may lead to too many resources being committed to the project, or, during contract bidding, result in not winning the contract, which can lead to loss of jobs. Accurate cost estimation is significant as it can help to classify and prioritize development projects with respect to an overall business plan. It can be used to determine what resources to commit to the project and how well these resources will be used. It can be used to assess the impact of changes and support re-planning. Projects can be easier to manage and control when resources are better matched to real needs. Customers expect actual development costs to be in line with estimated costs.

This main purpose of this paper is to review some of the machine learning techniques used in software cost estimation, summarise in a tabular form showing the various ML techniques used by most researchers, the datasets, evaluation metrics used and their limitations respectively. The full paper is organized in three sections, section 1 gives an introduction to the software project management and software cost and effort estimation, Section 2 reviews the machine learning SCE estimation techniques while Section 3 is conclusion and future work.

II. THE REVIEWED MACHINE LEARNING SCE ESTIMATION TECHNIQUES

According to [2], the software project cost estimation idea was conceived in 1960s and since then, so many cost estimation models have evolved by several researchers to determine the software project cost estimation. There are three broadly classified methods into which these models are categorized and they are; Algorithmic methods, Non-algorithmic methods and the Machine Learning methods.

[3] added that the algorithmic method was developed to deliver some mathematical equations to perform software cost estimation. The mathematical equations were based on historical data and research and made use of Source Lines of Code (SLOC), number of functions to carry out, cost drivers like language, design methodology, skill-levels and risk assessment and so on as inputs to the model. The algorithmic methods comprised the COCOMO, Function Point Analysis (FPA), Putnam model/SLIM respectively'. The non-algorithmic models on the other hand also known as non- parametric models make their estimation process based on deduction and analogy. They require the knowledge of previously completed software projects that are similar to the current software project for their estimation process. Examples include; expert judgment, Analogy based, Price to Win, Top-Down and Bottom-up models respectively.

An extensive review of some Authors work was done and the findings and analysis are summarised in Table I. It can be inferred from the Table 1 that most researchers in SCE used machine learning techniques to develop their models that provided accurate predictions and estimations. The datasets used mostly for the model predictions is the COCOMO datasets followed by that of NASA datasets. The evaluation metrics used by most researchers to evaluate the machine learning SCE developed models are the Mean Magnitude of Relative Error (MMRE), Percentage Relative Error Deviation (PRED) and Root Mean Squared Error (RMSE) respectively. Meanwhile, there were a lot of limitations and recommendations registered by most Authors reviewed for the accurate software project cost/effort prediction but the most common one is the recommendation of a hybrid machine learning model to improve the accuracy problem inherent in most SCE models.

Table 1: Machine Learning Software Cost Estimation Techniques Analysis

S/No	Author	Datasets Used	MLT Used	Evaluation Criteria	Accuracy/Improvement	Limitations/Recommendations
1	[4]	Software Analytics for Mobile App (SAMOA) Dataset	Multiple linear regressions (MLR), Multi-Layer Perceptron Neural Network (MLP-NN), Genetic Algorithm (GA) and Naïve Bayes forecasting approach (NBF)	MMRE, MRE, PRED(25)	Gave best accuracy Genetic Algorithm and suitable for mobile app effort estimation using SAMOA dataset.	(i) The selection of input data affected the accuracy of prediction. (ii) Lack of calibrated model/method to administer the scope of effort estimation for mobile apps.
2	[5]	Version 8 of ISBSG datasets, OWN Data Set for Agile	Genetic Algorithm (GA) and Artificial Neural Network (ANN)	PRED, and MMR. MMRE	The ANN gave a better accuracy than Genetic Algorithm	(i) Most old datasets of software engineering employed the waterfall methodology for software development (ii) Training and testing of the ML methods is not done with adequate numbers on this new (Agile) methodology. (iii) Software developers employ 4^{th} and 5^{th} generation languages and 1^{st} , 2^{nd} , and 3^{rd} generation languages have gone out of practice.
3	[6]	COCOMO 81	Genetic Programming, Neural Networks and Genetic Algorithms	MARE and MMRE	Proposed method outperformed most existing methods in estimation	N/A
4	[7]	International Software Benchmarking Standards Group (ISBSG) dataset	Decision Tree (DT), SGB, RF and SVR Kernel	Mann-Whitney U test, PRED (x), RMSE, (MMER), (MMRE), (MAE)	SVR RBF kernel based effort estimation technique yields better performance over other techniques	i) Did not apply ensembles of machine learning techniques for the software development effort estimation ii) To adopt Software Non-functional Assessment Procedure (SNAP) point approach for SCE iii) Application of various ML techniques over the SNAP point dataset to improve effort estimation accuracy.

						iv) New approaches based on AOP and FOP concepts to be identified in software estimation.
5	[8]	Desharnais dataset and COCOMO NASA dataset.	Fuzzy Logic Algorithm	MMRE and PRED	N/A	N/A
6	[9]	COCOMO 81, Albrecht, Desharnais, Maxwell ISBSG, China	Used GA and NN for feature selection; GA and imperial competitive algorithm (ICA) are used for clustering and MLP neural network used for modeling.	MMRE, MdMRE and PRED (0.25)	Results had it that the proposed model outperformed all the other methods for all the datasets and regression-based methods come next to it	To adopt the use of regression methods in modeling, fuzzy methods in clustering and using new sets of data with a larger number of records.
7	[10]	ISBSG dataset	Support Vector Machines, Neural Networks and Generalized Linear Models (SVM, MLP, GLM)	Mean Magnitude relative error (MMRE) and percentage relative error deviation (PRED).	Generated remarkable and accurate predictions due to the applied smart data preprocessing, three effective ML algorithms and cross-validation method for preventing overfitting occurrence	i) Addressed the impact of software sizing especially on effort estimation ii) To explore new trends in business analytics, such as prescriptive analytics. (iii) The impact of various different Feature Selection Methods to be considered for accuracy of machine learning effort and duration estimation models.
8	[11]	COCOMO dataset	Multilayer neural network technique using perception algorithm	Magnitude of Relative Error (MRE)	Result generated from multilayer neural system was then compared with Kaushik et al., (2013) and had a better accuracy and improvement.	To focus on neurofuzzy approach to software cost estimation
9	[12]	COCOMO dataset NASA dataset artificial dataset involves 100 different projects information from the COCOMO and NASA dataset	soft computing approach, Artificial Neural Networks,	Mean Magnitude of Relative Error (MMRE) and Pred(L)	The result indicated a 9.28% improvement in case of the accuracy of the estimation with the propose model.	To use soft computing techniques for software cost estimation
10	[1]	DESHARNAIS dataset from PROMISE Software Engineering Repository.	ANFIS, MATLAB Fuzzy Logic Gaussian membership	MAE, Correlation Coefficient and RMSE.	ANFIS technique has been able to outperform the Linear Regression Model in terms of the various performance criteria. The ANFIS hybrid model gave a better result	Its performance was not benchmarked with other hybrid methods of SCE but only the Linear Regression Model and thus advised to be done
11	[13]	Albrecht, China, Desharnais, Kemerer, Kitchenham, Maxwell, and Cocomo81	Ensemble techniques of averaging, weighted averaging, bagging, boosting, and stacking. stacking using a generalized linear model, stacking using decision tree, stacking using a support vector machine, and stacking using random forest	Mean Absolute Error, Root Mean Squared Error, and R^2.	stacking using random forest provides the best results compared with single model approaches	Recommended a hybrid model
12	[14]	NASA software projects dataset	Cuckoo Search (CS) algorithm	MMRE and PRED	Results obtained were seen to be better estimating model in comparison to other recent estimation models with respect to MMRE and PRED	To incorporate multiple evaluation estimation criteria to optimize the parameters and to investigate the suitability of the procedure for the accurate cost estimation.
13	[15]	COCOMO NASA 1, COCOMO NASA 2, COCOMO81 and Kaushik et al.2012	Random Forest (RF), Linear Regression (LR), Regression Tree (RT) and Support Vector Machine (SVM)	Correlation Accuracy, P-value, Significant Value, MMRE, A Measure, Rank Sum, Significant Value of Rank Sum	Indicated good performance and increase in the accuracy of Support Vector Machine for estimation of software project effort	To employ machine learning models that will perform ensemble stacking called blending, (ii) To ensemble the four machine learning models in order to optimize the predictive model. iii)To include percentage of accuracy to show the difference percentage between the predicted value and the actual effort value
14	[16]	60 NASA projects	K Nearest Neighbours algorithm (KNN) , Cascade Neural Networks (CNN) and Elman Neural Networks (ENN)	MMRE, RMSE, BRE	KNN is recommended to be used as a model based system for predicting the estimated cost of the projects to be built or developed	i) Non-computational methods can be used to find the software prediction and combine them with the computational methods ii) Different methods can be applied to solve the problem of predicting the cost closest to the real cost,

15	[17]	NASA	Naïve Bayes, Logistic Regression and Random Forests	AUC, CA, Precision and Recall, Confusion Matrix and ROC five folds cross-validation	The results of this work confirm the validity of data mining as an alternative for traditional estimation methods such as COCOMO.	Extending the research to cover more machine learning techniques and a hybrid of them
16	[18]	SEERA dataset	Naïve Bayes	Confusion Matrix, Accuracy	The accuracy of the future prediction system 86.59%. using the COCOMO II model the effort calculation accuracy is increased up to the 95.06% as compared to the SVM algorithm 93.45 %	To use hybrid machine algorithm to increase the accuracy of the future prediction of the profit and loss of the system more than 90% And overall effort calculation with the 97-98% accuracy.
17	[19]	COCOMO81, COCOMONASA, and COCOMONASA_2	Linear regression and K-Nearest Neighbors.	Correlation Coefficient, Mean squared error (MSE) and Mean magnitude relative error (MMRE)	percentage split and k-fold cross-validation	Algorithms like Decision Tree, Support Vector Machines, and Multi-layer Perceptron to be used to test the Effort Estimation with the existing datasets. Artificial Neural Network (ANN), Fuzzy Inference Systems (FIS) and Genetic Algorithms (GA) techniques for effort prediction
18	[20]	China dataset	Ensemble learning method bagging with base Learner Linear Regression, Support Vector Machine, Neural Network (MLP), MRules 5, REPTree, and Random Forest	Relative Absolute Error (RAE). Mean Magnitude of relative error (MMRE) (or mean absolute relative error). Root relative squared error (RRSE), Relative Absolute Error (RAE). Root relative squared error (RRSE)	Genetic Algorithm feature selection for the bagging M5 rule is the best method for predicting efforts with MMRE value 10%, and PRED (25), PRED (50) and PRED (75) have values 97%, 98% and 99%, respectively.	To use Ensemble Learning with different feature selection methods for predicting efforts estimation
19	[21]	NASA-93 and NASA-60	Dolphin algorithm and hybrid of dolphin and bat algorithm (DolBat).	Magnitude of Relative Error(MRE) and Mean Magnitude of Relative Error(MMRE) , PRED (25)	The dolphin swarm algorithm and hybrid bat algorithm (DolBat) were seen to optimize the cost estimation models COCOMO II coefficients (a, b)	To extend the work by the hybridization using a dolphin algorithm with other algorithms or using a new algorithm to estimate the efforts of software applications
20	[22]	Nasa 93, Nasa 60, COCOMO81 and Deshnaris	Expectation maximization (EM) algorithm for clustering. Fuzzy analogy with firefly algorithm	Mean magnitude of relative error (MMRE), and prediction (PRED)	The fusion of fuzzy and firefly algorithm on clustered datasets is exploited to improve the accuracy of the estimation.	To execute other implementations through other outstanding optimization processes

III. RESULTS

The results obtained from some of the literatures reviewed are illustrated in Figures 1, 2 and 3 respectively. From Figure I, it can be seen that the most commonly used datasets in software cost estimation/prediction using machine learning by researchers are the COCOMO and NASA datasets respectively. This is because they were the first datasets developed and used mainly Lines of Codes (LOC) of 1st, 2nd and 3rd generation programming languages as their major attributes. The datasets are also developed using the waterfall methodology in software development. SEERA and SAMOA datasets from the figure indicated that they were the least used datasets by researchers. The datasets were developed to handle flexibility and agile methodology of software development as could be seen in the mobile applications and other applications that evolve drastically due to technological advancement in the software industry.

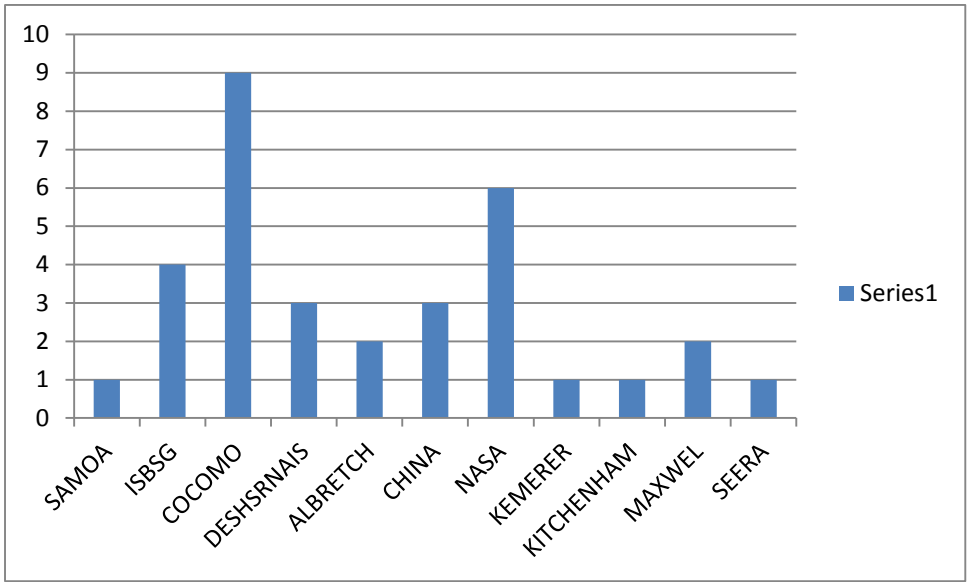

Figure 1: The Frequency of Datasets used in SCE by Authors Reviewed

Most of the models developed in software cost estimations using machine learning algorithms are usually evaluated to check their performance levels to ascertain if there is an improvement in the accuracy levels by the models or not. From the literature reviewed, it was noticed that amongst all the commonly used model performance metrics, the Mean Magnitude of Relative Error (MMRE), Percentage Relative Error Deviation (PRED) and Root Mean Squared Error (RMSE) are the most commonly used model evaluation performance criteria by researchers in SCE as seen in Figure 2. Therefore, researchers intending to perform further studies on machine learning models of SCE will find this knowledge useful for their model performance evaluation.

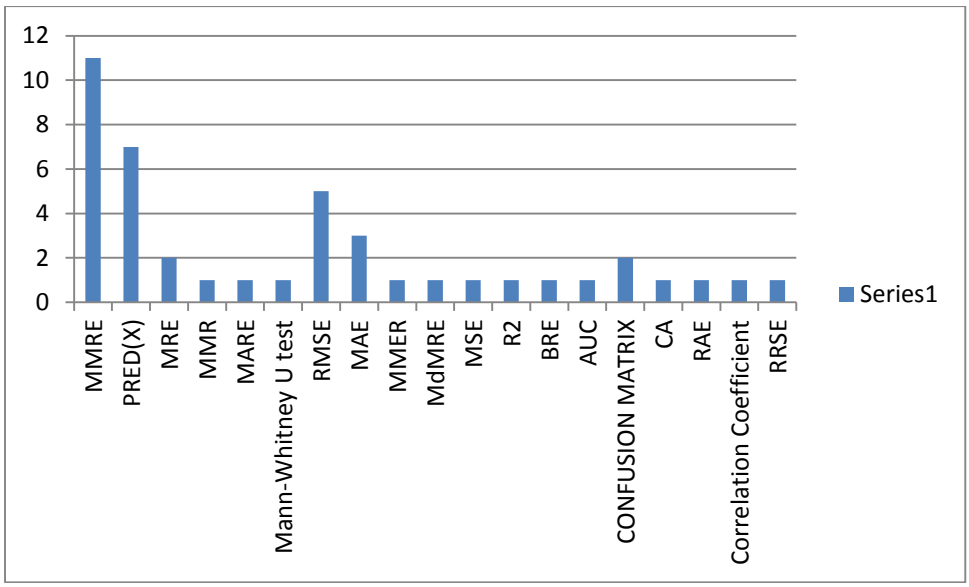

Figure 2: The Frequency of Model Performnce Evaluation Metrics used in SCE by Authors Reviewed

Figure 3 illustrates the result of the most frequently used machine learning algorithms in software project estimation by the Authors reviewed from Table 1.

It can be seen that Support Vector Machine (SVM), Multi-Layer Percetpron Neural Network (MLP-NN) and Artificial Neural Network (ANN) emerged the top most employed ML algorithms by resaechers. This is attributed to its simplicity, compatibility with most SCE datasets and high accuracy levels with minimal errors achieved by the use of these ML algorithms in developing and deploying the models.

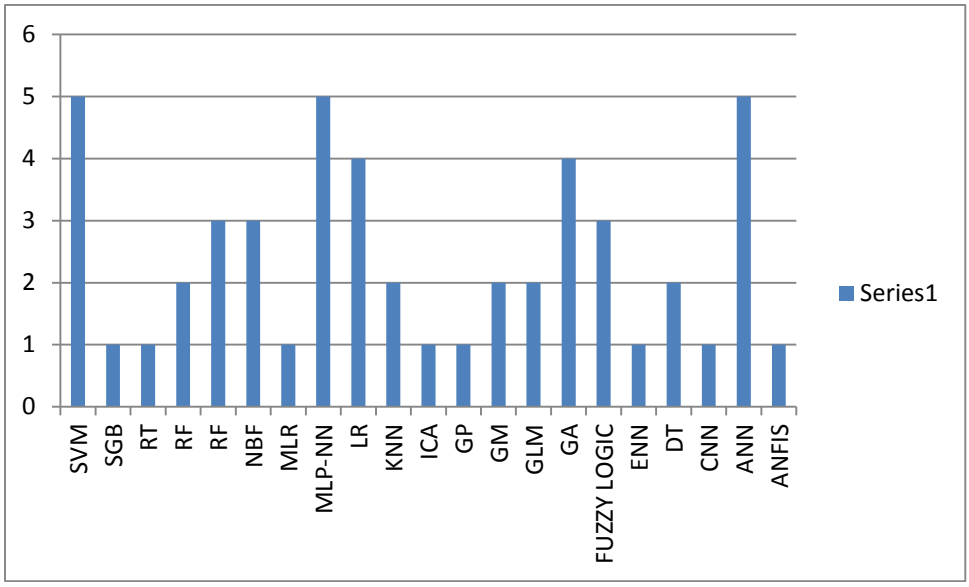

Figure 3: The Frequency of Machine Learning Algorithms used in SCE by the Authors Reviewed

V. CONCLUSION

This paper carried out an extensive literature review of the machine learning techniques used in SCE and arrived at an analysis showing results of the most commonly used datasets, model performnce evaluation metrics and the machine learning algorithms predominantly used by researchers in the field of software engineering. A brief introduction of the concept of sotware project cost estimation was done as well as the SCE techniques using machine learning. Tthe reseacrhes os some Authors were reviewed and summarised in Table 1. Barcharts were used to analyse the results indicating the most commonly Datasets used, Model Performance Metrics/criteria and the Machine Learning Algorithms employed by the Authhors of reseaches reviewed. The COCOMO and NASA datasets were seen to be the most widely used in SCE as well as the MMRE, PRED(x) and RMSE for the SCE models performance evaluation metrics/criteria. SVM, ANN and MLP-NN are the most widely used ML algorithms in developing the models required for SCE.

As a future direction, more reviews should be done to cover more authors amf their findings summarised and analysed to increase the scope of this work. Also, in the field of SCE, further work should be done to analyse the non-machine learning techniques used in SCE and compare with the scope of this work to foster a better understanding and direction for reseachers in the field of software engineering and SCE.

REFERENCES

[1]. Shukla, H. K., Singh, S. V. and Singh, R. B. (2021). Cost Estimation of Software by ANFIS based Artificial Intelligence Approach. International Journal of Research and Development in Applied Science and Engineering (IJRDASE) Available online at: www.ijrdase.com

[2]. Iqbal, S. Z., Idrees, M., Sana, A. B. and Khan, N. (2017). Comparative Analysis of Common Software Cost Estimation Modelling Techniques. Mathematical Modelling and Applications. Vol. 2, No. 3, 2017, pp. 33-39. doi: 10.11648/j.mma.20170203.12

[3]. Chirra, S.M.R. and Reza, H. (2019) A Survey on Software Cost Estimation Techniques. Journal of Software Engineering and Applications, 12, 226-248. https://doi.org/10.4236/jsea.2019.126014.

[4]. Pandey, M., Litoriya, R. and Pandey, P. (2019). Validation of Existing Software Effort Estimation Techniques in Context with Mobile Software Applications. Wireless Personal Communications https://doi.org/10.1007/s11277-019-06805-0. © Springer Science+Business Media, LLC, part of Springer Nature 2019

[5]. Tayyab, M. R., Usman, M. and Ahmad, W. (2018). A Machine Learning Based Model for Software Cost Estimation. © Springer International Publishing AG 2018 Y. Bi et al. (eds.), Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016, Lecture Notes in Networks and Systems 16, DOI 10.1007/978-3-319-56991-8_30.

[6]. Nawaz, M. A., Kumar,M., Zaki, M. D., Sunil, K. G., Prithvi, R., Neeraj, K. and Raja, B. D. (2020). A Methodology for Software Cost Estimation Using Machine Learning Techniques. International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020) (May 27, 2020)

[7]. Shashank, M. S. (2016). Effort Estimation Methods in Software Development using Machine Learning Algorithm. PhD Thesis submitted to the Department of Computer Science and Engineering National Institute of Technology Rourkela

[8]. Singh, S. P., Ch, S. S. (2020). A Novel Fuzzy Model for Software Cost Estimation. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4

[9]. Khazaiepoor, M., Bardsiri, A. K. and Keynia, F. (2019). A Dataset-Independent Model for Estimating Software Development Effort Using Soft Computing Techniques. Sciendo Journal of Applied Computer Systems. ISSN 2255-8691 (online) ISSN 2255-8683 (print) December 2019, vol. 24, no. 2, pp. 82–93 https://doi.org/10.2478/acss-2019-0011 https://content.sciendo.com

[10]. Pospieszny, P., Czarnacka-Chrobot, B. and Kobyliński, A. (2017). An Effective Approach for Software Project Effort and Duration Estimation with Machine Learning Algorithms, The Journal of Systems & Software . doi: 10.1016/j.jss.2017.11.066.

[11]. Arora, S. and Mishra, N. (2018). Software Cost Estimation Using Artificial Neural Network. Soft Computing: Theories and Applications, Advances in Intelligent Systems and Computing 584, Springer Nature Singapore Pte Ltd. M. Pant et al. (eds.), https://doi.org/10.1007/978-981-10-5699-4_6

[12]. Attarzadeh, I and Ow, S. H. (2011). Software Development Cost and Time Forecasting Using a High Performance Artificial Neural Network Model. R. Chen (Ed.): ICICIS 2011, Part I, CCIS 134, pp. 18–26, 2011. © Springer-Verlag Berlin Heidelberg 2011.

[13]. Varshini, P. A. G., Kumari, A. K. and Varadarajan, V. (2021). Estimating Software Development Efforts Using a Random Forest-Based Stacked Ensemble Approach. Electronics 2021, 10, 1195. https://doi.org/10.3390/ electronics10101195

[14]. Kumari, S. and Pushkar, S. (2017). Software Cost Estimation Using Cuckoo Search. Advances in Computational Intelligence, Advances in Intelligent Systems and Computing 509, DOI 10.1007/978-981-10-2525-9_17

[15]. Zakaria, N. A., Ismail, A. R., Ali, A.Y., Khalid, N. H. M. and Abidin, N. Z. (2021). Software Project Estimation with Machine Learning. International Journal of Advanced Computer Science and Applications, (IJACSA) Vol. 12, No. 6.

[16]. Abdulmajeed, A.A., Al-jawaherry, M. A. and Tawfeeq, T. M. (2021). Predict the Required Cost to Develop Software Engineering Projects by Using Machine Learning. Ashraf Journal of Physics: Conference Series 1897 012029.

[17]. BaniMustafa, A. (2018) Predicting Software Effort Estimation Using Machine Learning Techniques. 8th International Conference on Computer Science and Information Technology (CSIT) . 978-1-5386-4152-1/18/$31.00 ©2018 IEEE . ISBN: 978-1-5386-4152-1 . American University of Madaba Kings Highway, Madaba, Jordan, +96253294444, a.banimustafa@aum.edu.j

[18]. Bushra, Q. and Kadam, A. (2021). An Improved Technique for Software Cost Estimations in Agile Software Development using Soft Computing Techniques. IT (Information Technology) in Industry, Vol. 9, No.2. ISSN (Online): 2203-1731

[19]. Marapelli, B. (2019). Software Development Effort Duration and Cost Estimation using Linear Regression and K-Nearest Neighbors Machine Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2,

[20]. Khan, B., Naseem, R., Binsawad, M., Khan, M. and Ahmad, A. (2020). Software Cost Estimation Using Flower Pollination Algorithm. Journal of Internet Technology Volume 21 No.5. DOI: 10.3966/160792642020092105002

[21]. Fadhil, A.A., Alsarraj, R. G. and Altaie, A. M. (2020) Software Cost Estimation Based on Dolphin Algorithm. IEEE Acess Volume 8, 2020. Digital Object Identifier 10.1109/ACCESS.2020.2988867.

[22]. Resmi, V., Vijayalakshmi, S. R., Chandrabose, S. (2017). An effective software project effort estimation system using optimal firefly algorithm. Cluster Computing https://doi.org/10.1007/s10586-017-1388-0. © Springer Science+Business Media, LLC, part of Springer Nature 2017.