Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Analysis of Software Sizing and Project Estimation prediction by Machine Learning Classification

A. Sathesh[1], Yasir Babiker Hamdan[2]

[1]Department of Electronics and Communication Engineering, Eritrea Institute of Technology, Eritrea
[2]International University of Africa (IUA), Khartoum, Sudan

E-mail: [1]sathesh4you@gmail.com, [2]yasir20ap@iua.edu.sd

Abstract

In this study, the outcomes of trials with various projects are analyzed in detail. Estimators may decrease mistakes by combining several estimating strategies, which helps them maintain a close eye on the difference between their estimations and reality. An effort estimate is a method for estimating a model's correctness by calculating the total amount of effort needed. It's a major pain in the backside of software development. Several prediction methods have recently been created to find an appropriate estimate. The suggested SVM approach is utilized to reduce the estimation error for the project estimate to the lowest possible value. As a result, throughout the software sizing process, the ideal or exact forecast is achieved. Early in a model's development, the estimate is erroneous since the needs are not defined, but as the model evolves, it becomes more and more accurate. Because of this, it is critical to choose a precise estimate for each software model development. Observations and suggestions for further study of software sizing approaches are also included in the report.

Keywords: Software sizing, machine learning, project estimation, effort estimation, prediction techniques, Support Vector Machine (SVM)

1. Introduction

The endeavour of software sizing and project estimation have been necessary and demanding from the beginning of the computer age. As computing technology evolved, price of the hardware fell, but the price of the software rose. As a result, software development was less affected by inaccurate estimations. However, as hardware costs have decreased, personal

DOI: https://doi.org/10.36548/jucct.2021.4.006
Received: 03.01.2022, received in revised form: 10.02.2022, accepted: 19.02.2022, published: 28.02.2022

computers have been introduced, and software costs have increased, the effect of poor software estimations has been more prominent than ever before.

The software planners require attainable plans, including realistic estimates of timelines, size, resources, and risks, in order to achieve dependable software project success [1, 2]. The failure of software projects has become a common occurrence in the software engineering world. Many projects are shelved before they've even begun. The best place to begin is by addressing some of the most common causes of software project failure.

The software's size is estimated during the requirements phase of development. Sizing software is an important part of the software development process. Estimates of work effort and costs are accurate when they are based on accurate measurements [3]. Software managers may find it difficult to provide realistic deadlines if they don't have a clear idea of the extent of their functions. There are a variety of approaches and procedures for scaling software. Once the size of a piece of software has been known, the required effort to build that piece of software may be estimated [4-6].

1.1 Motivation of the research

Research Question:

Can lower project failure probability improve the machine learning-based model's accuracy prediction using feature transformation and feature selection?

In this age of automation, estimating software is a task best performed by hand, with just the most basics of tools. It became a challenge and a motivation to experiment with automating some or all of the estimating process. Gathering expert information in some form or another to aid the estimation process was another driving factor in the design and implementation of automated estimating approaches. According to the above discussed concepts, the following are the primary causes of project failures and software issues:

1. Inaccurately sizing a software development project
2. A software project's inability to be properly sized
3. An incorrect evaluation of workforce numbers and abilities, as well as an inability to appropriately design an adequate software development and support environment.

4. A lack of well-defined software activity requirements

2. Organization of the Research

This complete study piece is divided into the following sections: Section 3 contains preliminaries on software sizing and project estimation. Section 4 describes the recommended process for estimating the cost of a project. Section 5 evaluates the suggested approach experimentally and compares it to the conventional procedure. Finally, section 6 discusses prospective future developments.

3. Preliminaries

There have been a slew of approaches created over the years to figure out the amount of work put forth to develop a new app. Some of the most commonly used estimation approaches include functional point analysis, expert opinion, and estimation by analogy.

Accurate project estimation is a requisite to successfully complete a project. At the beginning of the development phase, projects are budgeted and planned in terms of cost, effort, and time. In order to accomplish a software project on schedule and under budget without compromising the quality of the programme, it is critical to accurately estimate the software development effort required. An accurate cost estimate is crucial to the success of a building project, and it impacts the decision-making of stakeholders in a software project and their willingness to accept the proposal [7-12].

An estimation model's capability may be measured by its bias, stability, and accuracy. When comparing actual prices to predicted costs, measures of tendency, strength, and accuracy are concerned with determining the amount difference between the two averages. The most commonly used assessment criteria are statistics such as mean, standard deviation, and coefficient of variation [13].

Measurement of software metrics is critical for a variety of reasons, including estimating programming execution, monitoring and regulating software project executions, decreasing faults during software development, and determining the efficacy of the system [14]. Manual labour is required for the procedures outlined above. As an example, in functional point analysis, a substantive count is done manually, which demands a great deal of time,

competence in a certain field, and extensive information. There will be no comparable project at some point in the future to quantify the whole effort [15].

Many machine learning approaches have been considered in the past, and many are already in use producing excellent results. Using the old software effort performance criteria, is inaccurate and unsatisfying. As a consequence, a wide range of indicators and cost-estimating methods are developed. Good conceptual and theoretical foundation, and statistically substantial experimental confirmation are the two features missing from most of them.

4. Proposed Methodology

To stay up with the current environment, machine learning is crucial since the model continues to improve its performance via data or experience. Machine learning may help minimise human effort and errors. Figure 1 shows the proposed architecture for project estimation prediction.

4.1 Extraction of data

Data mining is the process of gaining an understanding of the data in order to look for interesting patterns that may contain useful information. Data mining has a proven track record in the corporate world, and more recently, in the realm of science. A wide range of multidisciplinary approaches are used in data mining, including statistical, machine learning, and pattern recognition methods. Data mining in software project prediction, has lately attracted a lot of attention because of the enormous amount of inaccuracy in conventional estimating approaches and the continual progress of machine learning algorithms that might assist in delivering a more accurate forecast [16-19].

4.2 Preparation Stage

It is possible to increase the accuracy of machine learning by using a pre-processing strategy that removes unnecessary and superfluous characteristics. In addition to lowering cardinality, these alternatives allow the addition of an optional feature based on a lack of interaction between attributes and categorization [20].

4.3 Construction of the Proposed Model

It's a supervised machine learning approach for categorising issues that was created by Vapnik. The kernel idea utilises the categorization limit to discover the units underneath. Based on the boundaries, it distinguishes between the data sets. It may have been employed in multiple categories, such as space points and map data. Linear SVM is utilised for multiclass classification jobs. Linear SVM delivers good precision. SVM is the fundamental classification technique for conducting classification jobs in multi-dimensional space to develop a hypermarket that classifies case studies into distinct groups. It offers two sorts of functions: regression and classification tasks. Besides, it can handle certain steady variables [21-23].

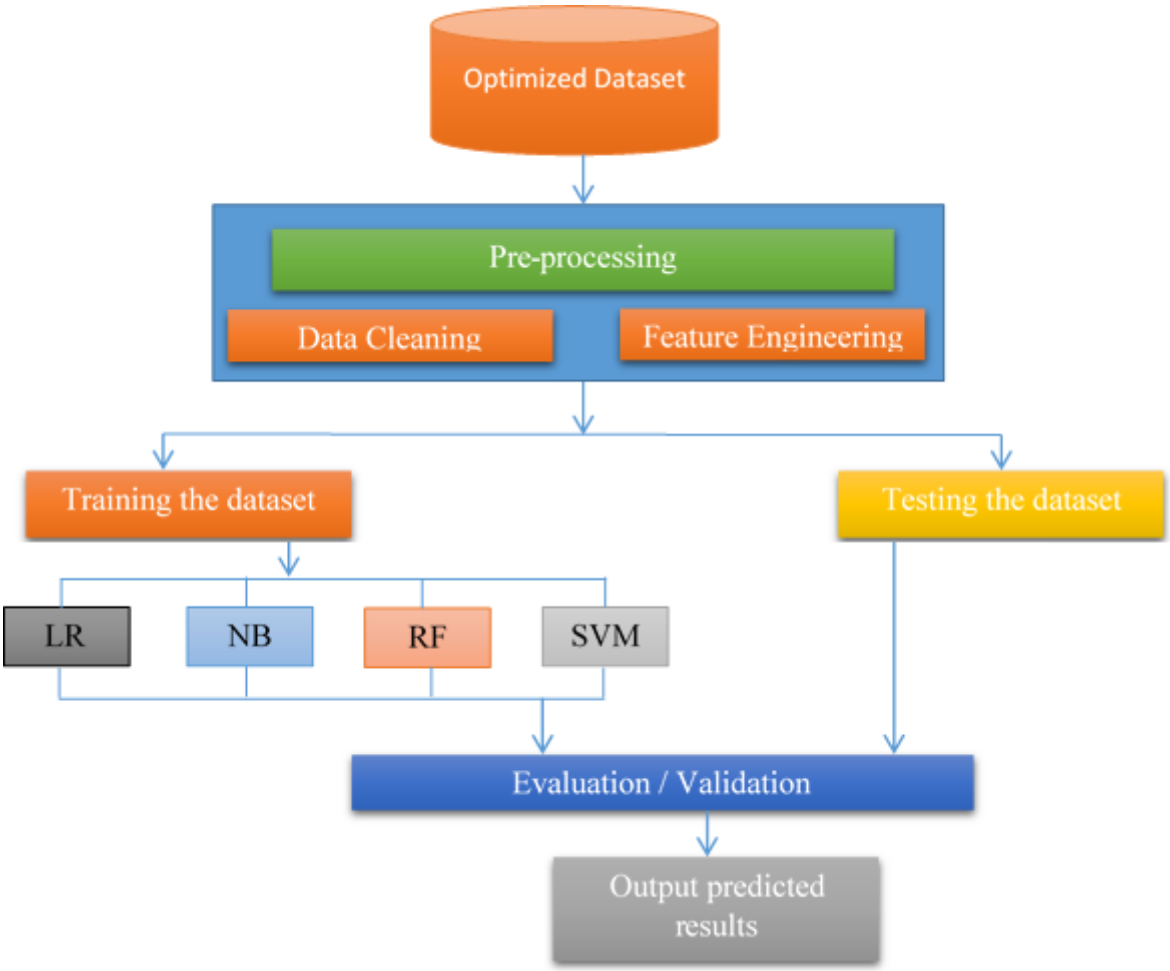

Figure 1. Proposed architecture for project estimation prediction

4.4 Parameters for Estimating Effort

The existing status of the effort estimate has been properly executed. Many evaluations and estimating techniques have been created for analysis in the last few years. But most of

them lack strong conceptualization, theoretical grounding, and statistically meaningful experimental confirmation. Moreover, most estimating measures were invented by people and implemented or validated in a minimum setting. In this work, it has been attempted to quickly describe several estimate metrics, including CC, MSE, RMSE and MAE. These are the prominent and widely utilised parameters for effort estimation [24, 25].

5. Results and Discussion

Desharnais dataset from PROMISE Software Engineering Repository, which has all 12 characteristics required to measure ML algorithm performance has been utilised in this experiment. One hundred and eleven individuals are included in the Desharnais dataset, with twelve characteristics. The dataset has been downloaded from:

https://www.kaggle.com/toniesteves/desharnais-dataset (in tabular form only).

Input data includes the evaluations of machine learning performance based on a 10-fold cross-validation test. The correlation coefficients suggest that the linear regression approach performs well in this comparison. A supervised attribute filter may be used in an experiment to increase the overall accuracy of the specified algorithms by selecting attributes from the provided dataset.

The Python programming language for the analysis of dataset, and the cross-validation technique for the complete procedure, have been utilized. The cross-validation approach has been used to create the most widely used machine learning algorithms, such as Linear Regression (LR), Multilayer Perceptron, Random Forest (RF) and Naïve Bayes (NB) algorithm. Table 1 summarises the performance metrics for each of these methods.

Table 1. Comparative analysis with performance metrics

S.No	Model	MSE	RMSE	MAE	CC	Accuracy
1	Linear Regression	0.354	0.5950	0.2155	0.6612	87.23%
2	Random Forest	0.7042	0.8392	0.2516	0.7128	89.03%
3	Naïve Bayes	0.1877	0.4322	0.3129	0.8007	82.56%
4	Proposed SVM	0.0353	0.1881	0.2019	0.5621	94.80%

Min-Max Accuracy has been used in order to get a sense of the closeness when looking at the average between the lowest and highest forecast. The greater the Min-Max accuracy setting, the more precise the results are. Figure 2 shows overall performance of error measures.

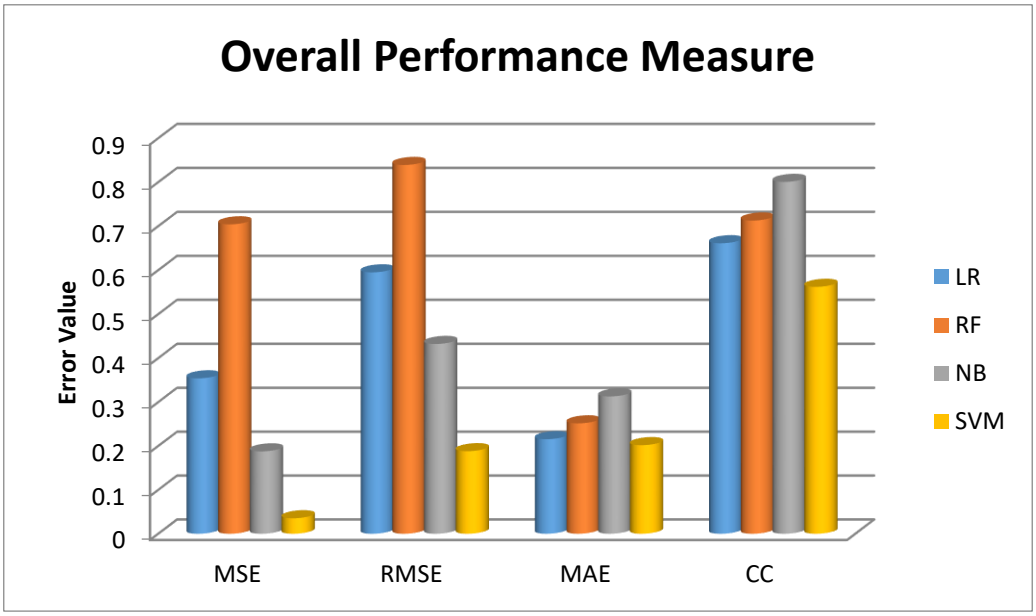

Figure 2. Overall Performances of Error Measures

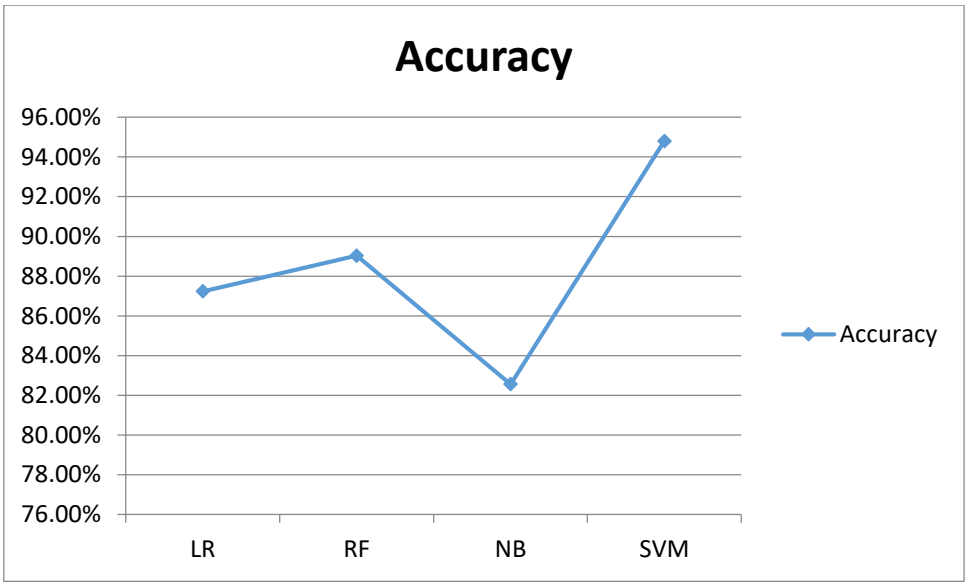

Figure 3. Accuracy measures for the proposed SVM

The correlation between the expected and actual data is used to determine accuracy. The degree of the correlation between the experiment's expected and actual values is measured using the Pearson product-moment correlation coefficient. When the correlation accuracy is high, the predicted and actual values move in the same direction. An experiment's importance

is determined by its P-Value, or computed probability. If the population correlation coefficient does not deviate from zero, then this project's null hypothesis is true. No substantial linear association exists between the population's control and experimental values. It's a nonparametric test that compares two samples to a single sample and determines whether the population rankings vary between the two samples. Enough data suggests that the sample as a whole does not include any individuals with identical distributions. Figure 3 shows the accuracy measures between the machine learning algorithms.

As each character's particular predictive capacity and degree of redundancy are taken into consideration, this method assesses the worth of the selection of qualities. This proposed algorithm with backtracking is used to look for a subset of characteristics in space.

6. Conclusion

The proposed SVM has been performed with higher accuracy results to predict the project estimation. To summarize, a number of current machine learning methods may be used to build prediction models. However, in order to correctly estimate the right and appropriate method, machine learning models must be in place. By including accuracy and other performance metrics, it is possible to demonstrate the difference between the expected and actual effort. As a result of sizing estimate inaccuracies, cost and schedule overruns are also likely to occur. In the future, it will be possible to assess and predict the relationships between underestimates of project size and subsequent overruns in effort, cost, and schedule. Moreover, Ensemble Stacking, also known as mixing, may be used to combine the four machine learning models in order to enhance the prediction model even more.

References

[1] M. Ruchika and J. Ankita, (2011). Software Effort Prediction using Statistical Machine Learning Methods. International Journal of Advanced Computer Science and Applications, vol. 2, no.1.

[2] Hamdan, Yasir Babiker. "Faultless Decision Making for False Information in Online: A Systematic Approach." Journal of Soft Computing Paradigm (JSCP) 2, no. 04 (2020): 226-235.

[3] Nayar, Nandini, Sachin Ahuja, and Shaily Jain. (2019). Swarm intelligence and data mining: a review of literature and applications in healthcare. Proceedings of the Third International Conference on Advanced Informatics for Computing Research.

[4] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110.

[5] Rekha Tripathi, Dr. P. K. Rai, (2016). Comparative Study of Software Cost Estimation Technique. International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 1.

[6] Sungheetha, Akey, and Rajesh Sharma. "Fuzzy Chaos Whale Optimization and BAT Integrated Algorithm for Parameter Estimation in Sewage Treatment." Journal of Soft Computing Paradigm (JSCP) 3, no. 01 (2021): 10-18.

[7] Nassif, A.B.; Capretz, L.F.; Ho, D.; Azzeh, M. A treeboost model for software effort estimation based on use case points. In Proceedings of the 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 15 December 2012; pp. 314–319.

[8] Andi, Hari Krishnan. "An Accurate Bitcoin Price Prediction using logistic regression with LSTM Machine Learning model." Journal of Soft Computing Paradigm 3, no. 3 (2021): 205-217.

[9] Amasaki, S.; Kawata, K.; Yokogawa, T. Improving cross-project defect prediction methods with data simplification. In Proceedings of the 2015 41st Euromicro Conference on Software Engineering and Advanced Applications, Madeira, Portugal, 28 August 2015; pp. 96–103.

[10] Karthigaikumar, P. "Industrial Quality Prediction System through Data Mining Algorithm." Journal of Electronics and Informatics 3, no. 2 (2021): 126-137.

[11] Wang, Y.-H.; Jia, J.; Qu, Y. The “Earth-Moon” model on software project risk management. In Proceedings of the 2010 International Conference on Machine Learning and Cybernetics, Qingdao, China, 14 July 2010; pp. 1999–2003.

[12] Kirubakaran, S. Stewart. "Study of Security Mechanisms to Create a Secure Cloud in a Virtual Environment with the Support of Cloud Service Providers." Journal of trends in Computer Science and Smart technology (TCSST) 2, no. 03 (2020): 148-154.

[13] Sigweni, B. Feature weighting for case-based reasoning software project effort estimation. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, London, UK, 13–14 May 2014; pp. 1–4.

[14] Hamdan, Yasir Babiker, and A. Sathesh. "Construction of Efficient Smart Voting Machine with Liveness Detection Module." Journal of Innovative Image Processing 3, no. 3 (2021): 255-268.

[15] Baolong, Y.; Hong, W.; Haodong, Z. Research and application of data management based on Data Management Maturity Model (DMM). In Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, 10 February 2018; pp. 157–160.

[16] Raj, Jennifer S. "Secure Data Sharing Platform for Portable Social Networks with Power Saving Operation." Journal of IoT in Social, Mobile, Analytics, and Cloud 3, no. 3 (2021): 250-262.

[17] Petkovic, D.; Sosnick-Pérez, M.; Huang, S.; Todtenhoefer, R.; Okada, K.; Arora, S.; Sreenivasen, R.; Flores, L.; Dubey, S. Setap: Software engineering teamwork assessment and prediction using machine learning. In Proceedings of the 2014 IEEE Frontiers in Education Conference (FIE) Proceedings, Madrid, Spain, 25 October 2014; pp. 1–8.

[18] Hamdan, Yasir Babiker. "Construction of Statistical SVM based Recognition Model for Handwritten Character Recognition." Journal of Information Technology 3, no. 02 (2021): 92-107.

[19] Azzeh, M.; Banitaan, S. An Application of Classification and Class Decomposition to Use Case Point Estimation Method. In Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 11 December 2015; pp. 1268–1271.

[20] Chen, Joy Iong Zong, and Joy Iong Zong. "Automatic Vehicle License Plate Detection using K-Means Clustering Algorithm and CNN." Journal of Electrical Engineering and Automation 3, no. 1 (2021): 15-23.

[21] Rajeshkanna, A., and K. Arunesh. "Optimizing Decision Tree Classification Algorithm with Kernel Density Estimation." In Innovative Data Communication Technologies and Application, pp. 257-266. Springer, Singapore, 2021.

[22] Pamina, J., J. Beschi Raja, S. Sam Peter, S. Soundarya, S. Sathya Bama, and M. S. Sruthi. "Inferring Machine Learning Based Parameter Estimation for Telecom Churn Prediction." In International Conference On Computational Vision and Bio Inspired Computing, pp. 257-267. Springer, Cham, 2019.

[23] Pratibha, C., K. Manish Reddy, L. Bharathi, M. Manasa, and R. Gandhiraj. "Simulation of Dual Polarization Radar for Rainfall Parameter and Drop Size Distribution

Estimation." In International Conference on Intelligent Computing, Information and Control Systems, pp. 424-433. Springer, Cham, 2019.

[24] Lingwal, Yogesh, Fateyh Bahadur Singh, and B. N. Ramakrishna. "Estimation of Differential code Bias and Local Ionospheric Mapping using GPS observations." In Proceedings of International Conference on Intelligent Computing, Information and Control Systems, pp. 809-824. Springer, Singapore, 2021

[25] Subramanian, R. Siva, and D. Prabha. "Optimizıng Naive Bayes Probability Estimation in Customer Analysis Using Hybrid Variable Selection." In Computer Networks and Inventive Communication Technologies, pp. 595-612. Springer, Singapore, 2021.

Author's biography

A. Sathesh completed his master's degree in the year 2006 and has published several papers in national and international journals. His areas of interest include wavelets and multi-resolution transforms for image denoising. Currently, he is occupying an academic position in Eritrea after having worked in a reputed University in South India for the past 5 years. He is pursuing his research work in the area of complex wavelets for image approximations with a deep learning approach.

Yasir Babiker Hamdan is presently working in International University of Africa (IUA), Khartoum, Sudan. His research is mainly focused on computer graphics, image processing, data mining, neural network algorithms and blockchain technologies.