Specify of Estimation Using C4.5 Algorithm of Software Project Estimation at the Point of Sales Application with Cocomo II

Kadinar Novel
Diploma in Informatics
AMIK Bina Sarana Informatika (BSI)
Jakarta, Indonesia
kadinar.ked@bsi.ac.id

Sfenrianto Sfenrianto
Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management,
Bina Nusantara University,
Jakarta, Indonesia, 11480
sfenrianto@binus.edu

Windu Gata
Master of Computer Science - Postgraduate Programs
STMIK Nusa Mandiri
Jakarta, Indonesia
windu.gata@gmail.com

Kaman Nainggolan
Master of Computer Science - Postgraduate Programs
STMIK Nusa Mandiri
Jakarta, Indonesia
kaman@nusamandiri.ac.id

Mochamad Wahyudi
Master of Computer Science - Postgraduate Programs
STMIK Nusa Mandiri
Jakarta, Indonesia
wahyudi@nusamandiri.ac.id

***Abstract*—In software development, it is required an appropriate estimate. One of the most commonly used software project estimation models is Constructive Cost Model(COCOMO II). The model is often used to obtain accurate results in estimating important factors such as cost and human resources. However, to obtain the more accurate estimation results, this study proposes a C4.5 algorithm based on COCOMO II estimation results. In this study, software project estimates are used in the Point of Sales (POS) applications. Based on these data with COCOMO II method, it is estimated that the schedule, staff, and cost are also specifying estimation from the result of COCOMO II using a C4.5 algorithm. The accuracy of the estimation results is around 90% with Algorithm C4.5. The value can be used as a reference for the development of the next POS software project.**

Keywords—cocomo II, algorithm C4.5, decision tree, point of sales, software project estimation

I. INTRODUCTION

Currently, the software has been used to support various business activities. The software is used in offices, small industries, manufacturing, entertainments, and others. In the development of software projects, it is required a good estimate of costs, human resources, and development schedule.

Estimated software development can use the COCOMO II Model approach [1] [2]. This approach is useful for making decisions in developing software projects [2]. However, the estimates using COCOMO II are less accurate [3].

Therefore, to improve the accuracy of software projects, Models such as COCOMO II must use The C4.5 Algorithm [4]. It has been used to overcome the problem of software project accuracy using COCOMO II model [5] [6]. Previous studies have shown that a C4.5 algorithms are more effective in predicting software projects [6] [7].

This study will estimate project development in Point of Sales application based on COCOMO II using C4.5 approach. The main purpose of this study is to analyse the effectiveness of using a C4.5 algorithm to determinate estimation from the result of COCOMO II.

II. RELATED WORK

Software is a physical abstraction that allows us to talk to hardware machines [12]. The software estimate is a prediction resources, such as development costs and timelines for specific software projects in certain environments, using defined methods [8].One of the most common approaches to software estimation is the COCOMO II model.

The COCOMO II model has helped the company to estimate cost software development projects [3]. Software cost estimation is required to complete all of the resources work on the software project [1]. Resources estimates must be made at the beginning of the project for accuracy cost and requirements of other resources [9]. Therefore COCOMO II model needs to estimate the accuracy of costs and other resources before the development of software projects.

In addition, in their research on The Impact of CMMI Based Software Process Maturity on COCOMO II's Effort Estimation concludes an accurate cost estimation of software development is essential in budgeting, project planning, and effective project management control [14]. Different software cost estimation models have different inputs [13]. A study reviewed the cost estimation methodology concluded that the most important reason for the failure of software projects has been the subject of much research in the last decade. According to the results of some of the research presented in their paper, the root cause of software project failure is an improper estimate at the initial stage of the project [14]. Thus, introducing and focusing on estimation methods is very crucial to achieve accurate and reliable estimates.

978-1-5386-5813-0/18/$31.00 ©2018 IEEE

In the current study, most of the current estimation techniques have been systematically illustrated. There is no estimation method that can present the best estimate in all situations and each technique can fit within a specific project. It is important to understand the principal of each estimation method to choose the best. The main reason is that the performance of each estimation method depends on several parameters such as project complexity, project duration, staff skills, development methods and so on. Some evaluation metrics and actual estimation examples have been done only to illustrate the performance of estimation methods (eg. Cocomo) [14]. Efforts to improve the performance of existing methods and introducing new estimation methods based on today's software project requirements can be a future work in this field.

An analysis of the estimated cost of software development using COCOMO II method in Inagata Technosmith concluded that cost estimation obtained from the results of this research can provide recommendations for further projects. Estimated time to complete the project using COCOMO II method was 12 months. Estimated employee to work on this project was 4 people, the estimated cost in 1 month amounted to Rp 8,396,000. So the estimated total cost of this project was Rp 100,752,000 [17].

In research estimation of software creation cost using COCOMO II method in information system [18] reporting development activities system of software project, concluded that:

- By using COCOMO II method can calculate or estimate business or cost and schedule or duration of time of a software project.
- In the first software project, the size used is derived from the conversion of Unadjusted Function Point (UFP) to the Source Line of Code (SLOC).
- On software projects that have been done and are intended for the development of the size used are derived from manual source code line calculation or using the help of freeware.
- On the other hand, it takes an accuracy of resource estimation from software project development. Limitations of the COCOMO II method in the estimation accuracy require another approach. One approach that can be used is the C4.5 algorithm.

A study reviewed the cost estimation methodology concluded that the most important reason for the failure of software projects has been the subject of much research in the last decade. Some evaluation metrics and actual estimation examples have been presented in this paper only to illustrate the performance of estimation methods (eg COCOMO). Attempting to improve the performance of existing methods and introducing new estimation methods based on today's software project requirements can be a future work in this field [13].

On the other hand, it takes an accuracy of resource estimation from software project development. Limitations of the COCOMO II method in the estimation accuracy require another approach. One approach that can be used is the C4.5 algorithm.

Responding to this, in the next research it is crucial a test with Algorithm C4.5 to get the results on the accuracy calculation of estimation of the COCOMO II method if it is used in project needs. The line of code (LOC) data originated from the repository of software projects on POS applications.

The C4.5 algorithm is one of the most popular Decision Tree methods, and more easily understood for classification or prediction. It is built in the form of a decision tree based on criteria as a learning model of the data sample [14].

Build decision trees from a project software prediction is based on the similarity data to describe precision, recall, and accuracy [9]. Such decision trees help managers decide whether projects are worth developing.The prediction accuracy of C4.5 in the software project can improve the accuracy percentage exceeds 40%. [10] and 89% [11]. Then, for recall and accuracy prediction of software project using decision tree are 84%, precision is 72% [9].

III. RESEARCH METHODOLOGY

The research method is done starting count the line of code (LOC) data of POS project, counting process the LOC data by using the SLOC Metric 3.0 application, estimation analysis using COCOMO II method, and result estimation with C4.5algorithm. Figure 1 shows the stages of the research conducted to estimate software projects.

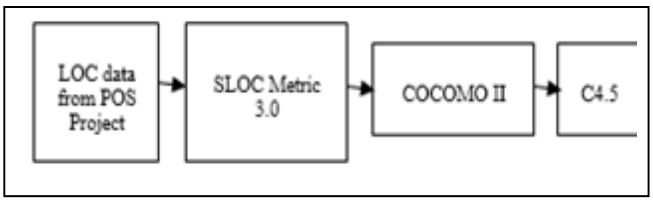

Fig. 1. The stages of the research

A. LOC Data and Sloc Matric 3.0

The data required for the software cost estimation analysis materials is based on the LOC data that have been collected. The data is derived from the software project repository in the POS application. The analysis using the COCOMO II method must be based on calculations from the SLOC that are obtained from the LOC data of POS application, ie. in the form of a file uses PHP programming language.

It is processed by counting the number of lines of code on by using the SLOC Metric 3.0 application that serve as the determination of the estimated effort of the POS software project. Figure 2 shown the results of SLOC data retrieval using SLOC Metric 3.0 from POS application.

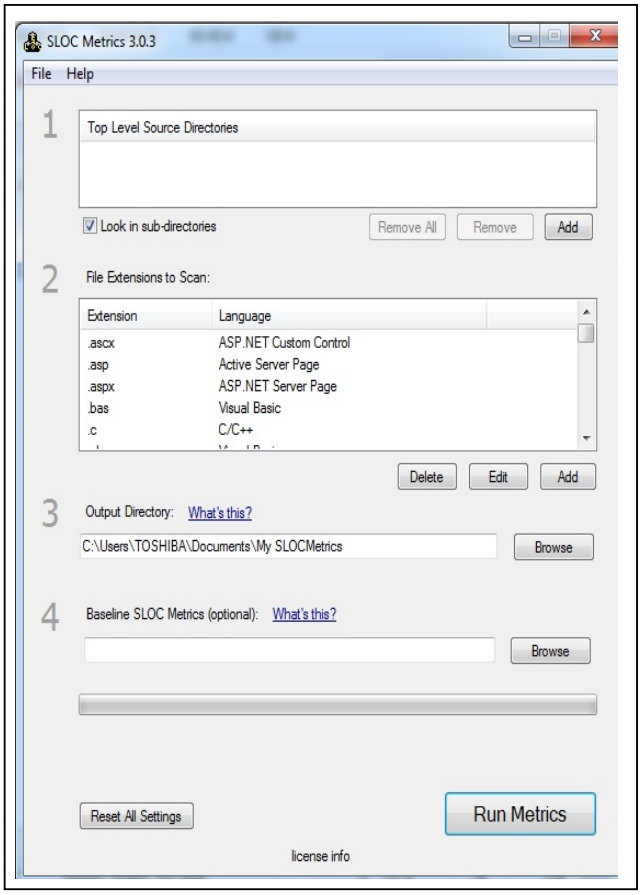

Fig. 2. SLOC Metric 3.0 Application

After all the files are inputted in SLOC Metric, then we obtained data from the SLOC or the modules of POS. There are several modules in the POS such as Sales, Order, Manual In Out Goods, Target, Returns, Stock, Mutation, Report, Customer, and Admin. Table I shows the number of files and SLOC data for the Costomer module based on PHP files.

TABLE I. SLOC DATA CUSTOMER MODULE

File	SLOC
customer.php	298
_print_transaction_cust.php	277
edit_customer.php	231
daily_cust.php	207
customer_transaction.php	191
detail_transaction_cust.php	187
add_customer.php	165
detail_pelanggan.php	158
_print_daily_all.php	152
_print_daily_customer.php	142
_print_customer.php	112
_print_daily_cust_total.php	100
_print_customer_extend.php	100
update_customer.php	95
history_customer.php	91
edit_point.php	82
_print_cust_point.php	63
insert_customer.php	43
update_point.php	22
get_poin.php	14
delete_customer.php	13
get_hp.php	9
get_email.php	9
Total	**2761**

TableII, shows the overall results of SLOC data from all modules POS application. Based on a total of 61275 SLOT data, the order module has the largest data, and the target module has the smallest data. This SLOT data will be used to estimation analysis using COCOMO II method.

TABLE II. THE RESULTS OF SLOC OF MODULES POS APPLICATION

Modules	SLOC
Sales	12092
Order	12431
Manual In Out Goods	7503
Target	1492
Retur	2801
Stock	3103
Mutation	2745
Report	8847
Customer	2761
Admin	7500
Total	**61275**

B. *Estimation Analysis using COCOMO II Method*

The estimation analysis using COCOMO II model required scale factor and cost driven for effort estimation [15]. The Scale Factor (SFi) used is PREC, FLEX, RESL, TEAM, and PMAT for each modul. The data obtained from the interview with users.

TABLE III. SCALE FACTOR

Module	*Scale Factor* SFi					∑SFi
	PREC	FLEX	RESL	TEAM	PMAT	
	Rating	Rating	Rating	Rating	Rating	
Sales	H	N	L	N	N	20.56
Order	H	N	L	N	N	20.56
Manual In Out Goods	H	N	L	N	N	20.56
arget	H	N	L	N	N	20.56
Retur	H	N	L	N	N	20.56
Stock	H	N	L	N	N	20.56
Mutation	H	N	L	N	N	20.56
Report	N	VH	L	L	N	20.86
Customer	H	N	L	N	N	20.56
Admin	N	H	L	N	N	20.79

TableIII, represents the Scale Factor value used to process estimation with COCOMO II, while the value obtained for PREC with H = 2.48 and N = 3.72 for FLEX with N = 3.04, VH = 1.01, for RESL with L = 5.62, for TEAM with value N = 3.29 and L = 4.38, for PMAT with value N = 4.68.

Then, the data cost driven in the estimation is required for each module as follows:

- ∏Emi Module of Sales = 16.74
- ∏Emi Module of Order = 16.74
- ∏Emi Module of Manual In Out Goods = 16.64
- ∏Emi Module of Target = 16.64
- ∏Emi Module of Retur = 16.64
- ∏Emi Module of Stock = 16.64
- ∏Emi Module of Mutation = 16.64
- ∏Emi Module of Report = 17.14
- ∏Emi Module of Customer = 16.64
- ∏Emi Module of Admin = 16.64.

Thus,the results of effort estimation for each POS moduleas follows:

- E Module of Sales = 31,2452
- E Module of Order = 32,2254
- E Module of Manual In Out Goods = 16,8857
- E Module of Target = 2,78588
- E Module of Return = 2,78588
- E Module of Stock = 6,30124
- E Module of Mutation = 5,49977
- E Module of Report = 31,9163
- E Module of Customer = 5,53554
- E Module of Admin = 16,9566

The above data is the result of estimation of effort from the calculation by COCOMO II method, which then will be processed to get estimation result of schedule, cost estimation, and personnel estimation. Table IV shows the calculation of estimated schedule (TDEV), personnel (P), cost (Cost).

TABLE IV. ESTIMATED OF SCHEDULE, COST, AND PERSONNEL

Module	TDEV(*in month*) = 3.67 x $(E)^F$	P = E / TDEV	Cost = Personil x *Avg Labor Cost*
Sales	9,757796	3,202071	16.010.353
Order	9,843809	3,273668	16.368.340
Manual In Out Goods	8,192545	2,061103	10.305.517
Target	4,910044	0,567385	2.836.924
Return	5,994966	0,938301	4.691.506
Stock	6,191453	1,017732	5.088.659
Mutation	5,956714	0,923289	4.616.444
Report	9,818934	3,250486	16.252.428
Customer	5,967697	0,927584	4.637.922
Admin	8,203367	2,067023	10.335.116

IV. RESULT

A. The result of cocomo II estimation

Based on the estimated schedule (TDEV), personnel (P), cost (Cost) of COCOMO II in table IV above are required for validation.It aims to determine the suitability of the results of the schedule, personnel, costs and project status. Table V shows the criteria for determining validation. A condition for determining the value of the attribute whether it will be worth No or Yes, and Priority or Not.

TABLE V. THE CRITERIA OF VALIDATION FOR SCHEDULE, PERSONEL, AND COST

No	Condition	Criteria
1	Project held (TDEV): < 6 month > 6 month	< 6 month : yes > 6 month : no
2	Personnel (P): < 2 staff > 2 staff	< 2 staff : yes > 2 staff : no
3	Cost (Cost): < Rp. 10,000,000 > Rp. 10,000,000	< Rp. 10,000,000 : yes > Rp. 10,000,000 : no

Data in Table VI shows estimation results based on criteria of schedule, personnel, and cost (see Table IV). It is used to define urgency requirement and project decision.

TABLE VI. THE RESULTS OF ESTIMATED OF SCHEDULE, COST, AND PERSONNEL BASED ON CRITERIA.

Module	Schedule Confirmity	Personnel Confirmity	Cost Confirmity	Urgency Requirment
Sales	No	No	No	Yes
Order	No	No	No	Yes
Manual In Out Goods	No	Yes	No	No
Target	Yes	Yes	Yes	No
Retur	Yes	Yes	Yes	No
Stock	No	Yes	Yes	No
Mutation	Yes	Yes	Yes	No
Report	No	No	No	Yes
Customer	Yes	Yes	Yes	No
Admin	No	No	No	No

Conformity is the schedule for the target, return, transfer and customers module (project held > 6 month). Personnel suitability is the in or out goods manual, target, return, stock, mutation and customer module (personnel > 2 staff). While the cost estimation results, there are five modules that have a cost match, namely: target, return, stock, mutation and customer (cost > Rp. 10.000.000). Thus, urgency requirement is the sales, order, transfer and report module.

B. The result of C4.5 estimation

The results of estimation schedule, cost, and personnel will be used to data set (data training sample). It can be processed in the C4.5 algorithm in addition attributes of conformity and ugency (Table V represents). The following are the steps of the C4.5 classification algorithm model [16]:

- Calculating the number of cases for priority rather than priority and entropy from all cases. Entropy total rows are calculated based on training data using the equation:

$$\text{Entropy}(i) = -\sum_{j=1}^{m} f(i,j).\text{Log2}\, f[(i,j)] \quad (1)$$

Entropy(total) = (-4/10 * *log2*(4/10)) + (-6/10 * *log2*(6/10))= 0.97095059445

- Then calculate the entropy and gain values of each attribute, for example below is calculation of the entropy and gain values for attributes Conformity Schedule:

$$\text{Entropy}(i) = -\sum_{j=1}^{m} f(i,j).\text{Log2} f[(i,j)] \quad (2)$$

$$\text{Yes} = \left(-\tfrac{1}{4} * log2\left(\tfrac{1}{4}\right)\right) + \left(-\tfrac{3}{4} * log2\left(\tfrac{3}{4}\right)\right)$$

$$\text{No} = \left(-\tfrac{3}{6} * log2\left(\tfrac{3}{6}\right)\right) + \left(-\tfrac{3}{6} * log2\left(\tfrac{3}{6}\right)\right)$$

$$\text{Gain} = Etotal - \sum_{j=1}^{m} total/totalKasus.E \quad (3)$$

$$\text{Gain} = 0.971 - \left(\frac{4}{10} * 0.811\right) + \left(\frac{6}{10} * 1\right)$$

$$= 0.9709 - 0.926$$

$$= 0.0469.$$

Entropy and gain calculations can be seen as follow:

1) Schedule Confirmity for Gain 0,046, Entropy Yes 0,811 and Entropy No 1
2) Personel Confirmity for Gain 0,046, Entropy Yes 1 and Entropy No 0,811
3) Cost Confirmity for Gain 0, Entropy Yes 0,971 and Entropy No 0,971
4) Urgency Confirmity for Gain 0,042, Entropy Yes 0 and Entropy No 918.

The data from Table VI above is a value for making a decision tree.Figure 3 shows the decision tree to determine the module priority of the POS application project. The main purpose of analyzing data using the decision tree algorithm is to get the rule which will be used for decision making on the module priority. Then Table VII shows the result of project decision for priority.

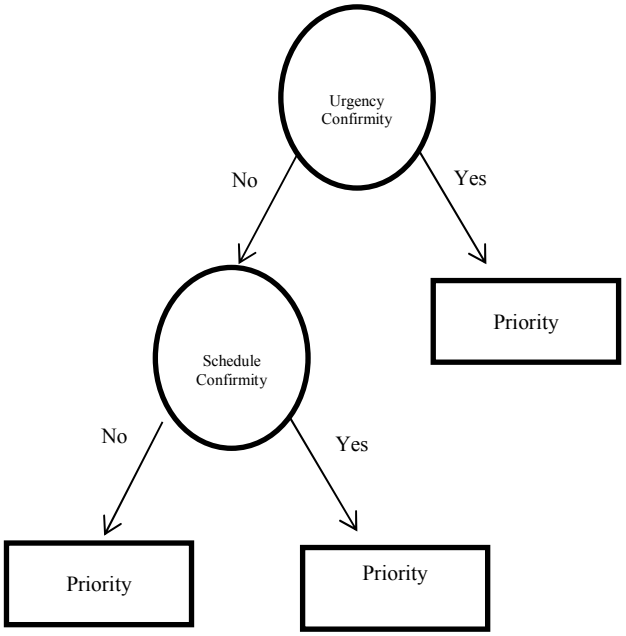

Fig. 3. Decision Tree Specify of Estimation

The rules derived from Figure 3 above are (1) If Urgency = "Yes", then "Priority"; (2) If Urgency = "No" and Schedule Conformity = "Yes", then "Priority"; and If Urgency = "No" and Schedule Conformity = "No", then "No". Tabel VII shows the result of project decision based on decision tree rules.

TABLE VII. THE RESULTS OF PROJECT DECISION BASED ON DECISION TREE RULES.

Module	Schedule Confirmity	Personnel Confirmity	Cost Confirmity	Urgency Requirment	Project Decision
Sales	No	No	No	Yes	Priority
Order	No	No	No	Yes	Priority
Manual In Out Goods	No	Yes	No	No	No
Target	Yes	Yes	Yes	No	Priority
Retur	Yes	Yes	Yes	No	Priority
Stock	No	Yes	Yes	No	No
Mutation	Yes	Yes	Yes	No	Priority
Report	No	No	No	Yes	Priority
Customer	Yes	Yes	Yes	No	Priority
Admin	No	No	No	No	No

Based on Table VII the project decision of priority is the sale, order, target, return, mutation, report, and customer module.

The result of the decision tree that has been done is required to measure the accuracy level using K-Fold Cross method (see Table VIII). Based on K-Fold Cross method using C4.5 Algorithm it produces accuracy equal to 90%.

TABLE VIII. K-FOLD CROSS VALIDATION

	True Prioritas	True No	Class Precision
Pred. Priority	6	1	85.71%
Pred. No	0	3	100.00%
Class recall	100.00%	75.00%	
accuracy: 90.00% +/- 30.00% (micro: 90.00%)			

CONCLUSIONS

Based on the results of POS modules estimation using COCOMO II the important conclusions of this study are:
(1) Conformity the schedule is the target, return, transfer and customers module. (2) Personnel suitability is the in out goods manual, target, return, stock, mutation and customer module. (3) There are five modules that have a cost match, namely: target, return, stock, mutation, and customer. (4) The urgency requirement is the sales, order, transfer and report module. While estimation result with the decision tree method of C4.5 Algorithm for the decision of priority is the sale, order, target, return, mutation, report, and customer module that produces accuracy equal to 90%. The results of the estimation can be used for subsequent project development.

The recomendations in the next research are: (1) For further research, in estimating software costs, it is desirable to use sub models other than Early Design in order to gain a better understanding of this COCOMO II method; (2) Further research can improve this research by adding other cost estimation methods as a comparison with the results of this study.

REFERENCES

[1] T. N. Sharma"Analysis of software cost estimation using COCOMO II." International Journal of Scientific & Engineering Research 2.6 (2011): pp. 1-5.

[2] J. Živadinović, Z. Medić, D. Maksimović, A. Damnjanović, & S. Vujčić Methods of effort estimation in software engineering. In *International Symposium Engineering Management and Competitiveness.* (2011, June), Zrenjanin, Serbia.

[3] R. Dillibabu, and K. Krishnaiah. "Cost estimation of a software product using COCOMO II. 2000 model–a case study."International Journal of Project Management, vol.23, no.4, pp. 297-307, 2005.

[4] X. Huang, D. Ho, J. Ren, and L. F. Capretz, L.F. Improving the COCOMO Model using a NeuroFuzzy Approach, Applied Soft Computing, vol. 7, no. 1, pp. 29-40, 2011.

[5] Q. Song, M. Shepperd, X. Chen, and J. Liu. Can KNN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation, The Journal of Systems and Software, vol. 81, pp. 2361-2370, 2008.

[6] F. S Gharehchopogh, I. Maleki, A. Kamalinia, and H. M. Zadeh, Artificial bee colony based constructive cost model for software cost estimation. *Journal of Scientific Research and Development*, Vo. *1, no.* 2, pp. 44-51, 2014.

[7] K. Mao, Y. Yang, M. Li, and M. Harman, Pricing crowdsourcing-based software development tasks. In *Proceedings of the 2013 international conference on Software engineering*, pp. 1205-1208, (2013, May, IEEE Press.

[8] M. Chemuturi, Software Estimation Best Practices, Tools & Techniques: A Complete Guide for Software Project Estimators, 2009, USA: J. Ross Publishing.

[9] F. S Gharehchopogh, Neural networks application in software cost estimation: a case study. In *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on* , pp. 69-73), IEEE.

[10] T. Zimmermann, N. Nagappan, H. Gall, E. Giger, and B. Murphy,. Cross-project defect prediction: a large scale experiment on data vs. domain vs. process. In *Proceedings of the the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, (2009, August) , pp. 91-100), ACM

[11] Sfenrianto, P. Indah, and R. Broer, Naïve Bayes Classifier Algorithm Particle Swam optimization for classification of Cross Selling (Case study: PT. TELKOM Jakarta), Cyber and IT Service Management, International Conference on IEEE, 2016.

[12] A. M. Langer *Analysis and Design of Information Systems* (3rded). 2008. London: Springer,.

[13] M. Al Yahya, R. Ahtnad, and S. Lee. "*Impact of CMMI Based Software Process Maturity on COCOMO II's Effort Estimation*". The International Arab Journal of Information Technology. vol. 7. No. 2. April 2010.

[14] Khatibi,Vahid andN. A. Jawawi, Dayang. "Software Cost Estimation Methods: A Review". Journal of Emerging Trends in Computing and Information Sciences. Volume 2 No. 1 2010. ISSN 2079-8407

[15] Boehm, B., 2001. Software Cost Estimation with COCOMO II.

[16] Nofriansyah, Dicky.Konsep Data Mining Vs Sistem Pendukung Keputusan. 2014, Yogyakarta: Deepublish.

[17] Damayanti, D. E. Suprapto, Kusuma, and A. R. Perdana. Analisis Estimasi Biaya Pembuatan Perangkat Lunak Menggunakan Metode COCOMO II di Inagata Technosmith (Studi Kasus : Sistem Informasi Monitoring dan Evaluasi Penerimaan Beasiswa Santri Berprestasi UIN Malang). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 1, No. 10, 2017 (e-ISSN: 2548-964X).

[18] Primaraka, A. Handoyo, E. Isnanto, and Rizal. "Estimasi Biaya Pembuatan Perangkat Lunak Menggunakan Metode Cocomo II Pada Sistem Informasi Pelaporan Kegiatan Pembangunan," (Project report), 2011.