TIME EFFORT PREDICTION OF AGILE SOFTWARE DEVELOPMENT USING MACHINE LEARNING TECHNIQUES

Muchamad Bachram Shidiq[1*], Windu Gata[2], Sigit Kurniawan[3], Dedi Dwi Saputra[4], Supriadi Panggabean[5]
Universitas Nusa Mandiri, Jakarta, Indonesia[1,2]
Universitas Teknologi Muhammadiyah Jakarta, Jakarta, Indonesia[3]
Universitas Siber Indonesia, Jakarta, Indonesia[4]
Universitas Darunnajah, Jakarta, Indonesia[5]
E-mail address: 14210136@nusamandiri.ac.id[1], windu@nusamandiri.ac.id[2],
sigit@utmj.ac.id[3], dedi.dwi.s@cyber-univ.ac.id[4], supriadipanggabean@darunnajah.ac.id[5]

Received: 17, Juni, 2023 Revised: 16, October, 2023 Accepted: 01, December, 2023

ABSTRACT

To run a software development project, an effective and efficient project management mechanism is needed to coordinate the activities carried out. The agile method was developed because there are several weaknesses in the classic method that can interfere with the course of the software development process according to user desires. However, in applying agile methods, time effort estimation cannot be done properly. This can cause project managers to have difficulty preparing resources in software development in scrum projects. For this reason, this research aims to predict the time effort of agile software development using Machine Learning techniques, namely the Decision Tree, Random Forest, Gradient Boosting, and AdaBoost algorithms, as well as the use of feature selection in the form of RRelieff and Principal Component Analysis (PCA) to improve prediction accuracy. The best-performing algorithm uses Gradient Boosting k-fold validation with PCA with an MSE value of 2.895, RMSE 1.701, MAE 0.898, and R2 0.951.

Keywords: AdaBoost, Agile, Decision Tree, Gradient Boosting, Random Forest.

1. INTRODUCTION

Nowadays, the use of Information Technology (IT) has a very important role in supporting the organization's business processes. Information Technology has a role as an important Business Enabler for organizations. Therefore, good governance is needed to ensure the continuity of business processes and get maximum benefits from IT implementation. IT management covers all aspects, from managing a reliable IT infrastructure to developing quality software (Peraturan Presiden Republik Indonesia, 2018). The quality of software development can be measured by three criteria, namely project time, project financing, and product results that meet expectations (Richards, 2007). There are several software development methods used, including waterfall, spiral, and agile. The agile method was developed because there are several weaknesses in the classic method that can interfere with the course of the software development process according to user wishes.

Pusat Sistem Informasi dan Teknologi Keuangan (Pusintek) is one of the echelon 2-level organizational units, under the Ministry of Finance of the Republic of Indonesia which, based on the Minister of Finance Regulation Number 118/PMK.01/2021, is tasked with developing software that will be used by stakeholders (Kementerian Keuangan RI, 2021). Since mid-2016, Pusintek has started implementing an agile development approach using the Scrum framework. Scrum is a methodology used in software development using incremental and iterative methods (Mahnic & Drnovscek, 2005). In determining software development projects, currently, this organization only uses estimation methods in the form of planning poker or expert estimation by considering the complexity of the software and the ability of organizational resources to complete the project on time and in accordance with the wishes of stakeholders. So that this has the potential to cause problems if, within a certain period of time, this organization gets an assignment to develop software with more than one project in parallel, which causes difficulties for project managers to allocate resources so that the delivery of software deliverables becomes late and / or not in accordance with the wishes of stakeholders.

Therefore, it is highly recommended to estimate using alternatives that do not have limitations as in the planning poker method, one of which uses Machine Learning techniques. Based on the above background, time effort prediction is one of the important points in the success of software development project management with agile methods, so this research will investigate the prediction of agile software development time effort using Machine Learning techniques conducted at the Pusintek Secretariat General of the Ministry of Finance.

This research will use several Machine Learning techniques, namely the Decision Tree algorithm, Random Forest, Gradient Boosting, and AdaBoost, as well as the use of feature selection in the form of Rrelieff and Principal Component Analysis (PCA) to improve prediction accuracy.

2. THEORY

The agile software development method is a software development approach that involves structured teamwork, responsiveness to change, and a focus on delivering a valuable product (Castillo, 2016). The origin of the agile methodology concept comes from the iterative and incremental approach to software development, where requirements and solutions will continue to evolve through an execution plan divided into iterative stages. At each stage, an evaluation of the output is conducted to ensure the project is on track. Agile methodologies in software development emphasize interaction and teamwork, rather than processes and tools. It is more important to produce well-functioning software than a complete document. Collaboration with clients or users is also favored over formal contract negotiations. In addition, being responsive to change is considered more important than just following an existing plan (Hohl et al., 2018).

There are several Machine Learning techniques that can be applied to Scrum to measure effort estimation. Here is an example of the algorithm and how it is used, as follows:

1. Decision Tree, Decision Tree Algorithm is a decision-making method used in machine learning. It is a graphical representation in the form of a decision tree, where each node in the tree represents a decision based on features or attributes (Rodríguez Sánchez, Vázquez Santacruz, & Cervantes Maceda, 2023).

2. Random Forest, Random Forest is a machine learning technique used to generate more accurate predictions by considering multiple decision trees. Random Forest algorithm is one of the ensemble learning methods in machine learning that utilizes a group of decision trees that work independently. Collectively, these decision trees form a "forest" that generates predictions by combining the results from each tree (Rodríguez Sánchez et al., 2023).
3. Gradient Boosting, The Gradient Boosting algorithm is an ensemble learning method that is also used for prediction in machine learning. Gradient Boosting builds a prediction model by combining several simpler models, in this case a Decision Tree (Nassif, Capretz, & Ho, 2012).
4. AdaBoost, AdaBoost (Adaptive Boosting) is one of the Machine Learning algorithms that belongs to the boosting algorithm category. The main goal of Adaboost is to improve the performance of a prediction model by combining several weak models (weak learner) into one stronger model (strong learner). Weak models in this context can be simple prediction models such as Decision Tree that have limited depth (Rodríguez Sánchez et al., 2023).

The process of choosing the most pertinent and instructive subset of features from the total features present in the dataset is known as feature selection. Some feature selection techniques that can be implemented are RRelieff feature selection and Principal Component Analysis (PCA) feature selection. Relief is an algorithm developed by (Kira & Rendell, 1992) that takes a filter method approach to feature selection that is highly sensitive to feature interactions.

Initially, the algorithm was designed to be applied to binary classification problems with discrete or numeric features (Kira & Rendell, 1992). Meanwhile, Feature Selection Principal Component Analysis (PCA) is a technique used to identify the most important or influential features in a dataset using the PCA method. PCA itself is a method used to reduce the dimensionality of a dataset by projecting the original variables into a lower principal component space (Khammas et al., 2015).

Evaluation and regression evaluation models that can be used in order to predict effort are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R2). MSE (Mean Squared Error) calculates the average of the squared difference between the model prediction and the true value in the dataset (Baura, 2002). Mathematically, the MSE formula is as follows:

$$MSE = \frac{\sum_{t=1}^{n}(At-Ft)^2}{n} \qquad (1)$$

where:
n is the number of data in the dataset
At is the actual value of the tth data
Ft is the predicted value of the tth data
RMSE (Root Mean Squared Error) calculates the root mean square of the difference between the predicted value and the true value in the dataset (Baura, 2002). Mathematically, the RMSE formula is as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(At-Ft)^2}{n}} \qquad (2)$$

where:
n is the number of data in the dataset
At is the actual value of the tth data
Ft is the predicted value of the tth data

MAE (Mean Absolute Error) measures the average of the absolute difference between the predicted and true values in the dataset (Baura, 2002). Mathematically, the MAE formula is as follows:

$$RMAE = \frac{1}{N}\sum_{t=1}^{n}|At - Ft| \qquad (3)$$

where:
n is the number of data in the dataset
At is the actual value of the tth data
Ft is the predicted value of the tth data
R2 (R-squared) measures how well the model fits the data, by comparing the difference between the actual value and the predicted value to the difference between the actual value and the average of the actual values (Chicco, Warrens, & Jurman, 2021). Mathematically, the R2 formula is as follows:

$$R2 = 1 - (SSres \,/\, SStot) \qquad (4)$$

where:
SSres is the sum of squares of the difference between the actual value and the predicted value (residual sum of squares).
SStot is the sum of squares of the difference between the actual value and the average actual value (total sum of squares)

3. METHOD

The research method is proposed using the Cross-Industry Standard Process for Data Mining (CRISP-DM). The method consists of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. As shown in figure 1.

3.1. Business Understanding

The Business Understanding stage conducts an understanding of the object of research. Understanding the object of research is done by taking research datasets. The dataset used is the scrum project dataset of the Service Travel Application (Perjadin Application) which is software developed using agile methods at Pusintek.

3.2. Data Understanding

This stage is the process of understanding the data that will be used as input material to the next stage, namely preprocessing. The dataset is a collection of historical backlog data which is translated into 165 tasks, 53 user stories which are divided into 7 sprints from 2021 to 2023. Figure 2 shows brief information on the dataset, namely the completion time of each sprint visualized using a bar plot.

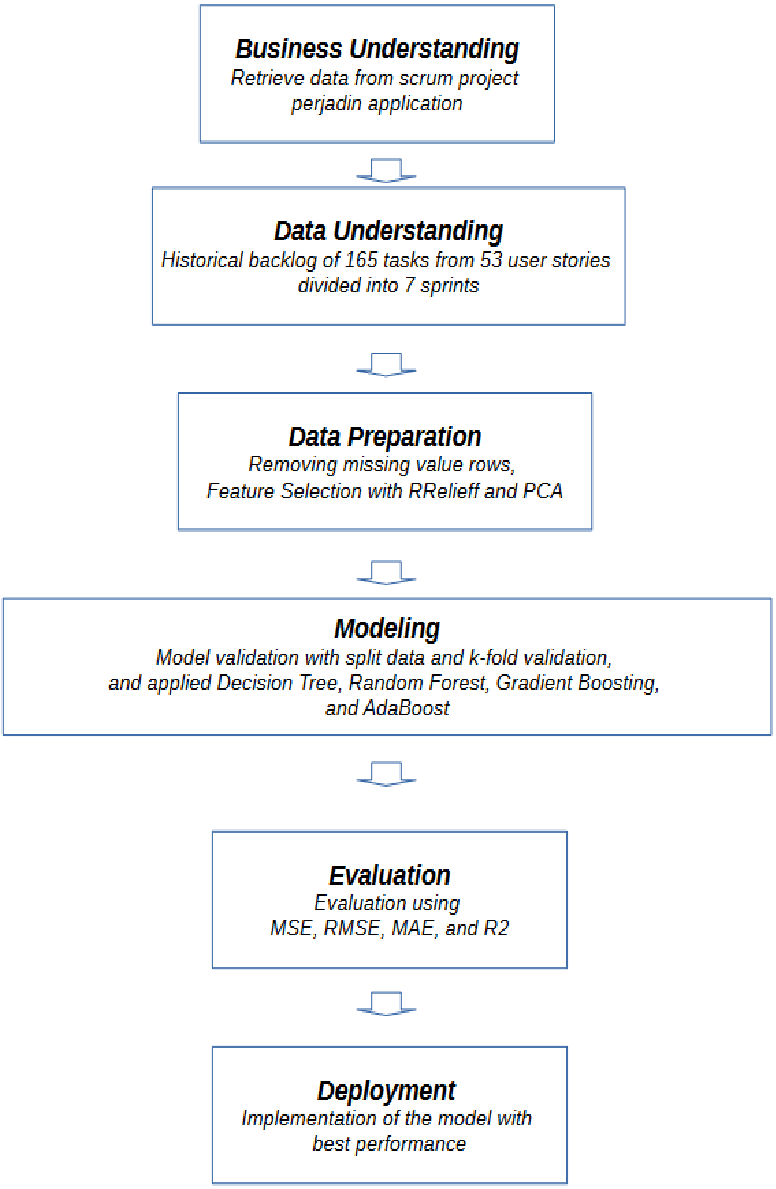
Business Understanding
Retrieve data from scrum project perjadin application
Data Understanding
Historical backlog of 165 tasks from 53 user stories divided into 7 sprints
Data Preparation
Removing missing value rows,
Feature Selection with RRelieff and PCA
Modeling
Model validation with split data and k-fold validation,
and applied Decision Tree, Random Forest, Gradient Boosting, and AdaBoost
Evaluation
Evaluation using
MSE, RMSE, MAE, and R2
Deployment
Implementation of the model with best performance

Figure 1. Proposed research model.

This stage also conducted an understanding to find methods with the best regression model approach in order to predict software development effort with agile methods. The algorithms used are Decision Tree, Random Forest, Gradient Boosting, and AdaBoost, which are carried out with split data training-testing and k-fold validation methods. And combined with feature selection RRelieff and Principal Component Analysis (PCA).

3.3. Data Preprocessing

After doing data understanding, it was found that there were 165 instances, there were some missing data detected. So it is necessary to do data preprocessing by running remove rows with missing values, so that 154 instances are obtained. In this study, feature selection will be carried out using 2 (two) different methods, namely the RRelieff method and Principal Component Analysis (PCA).

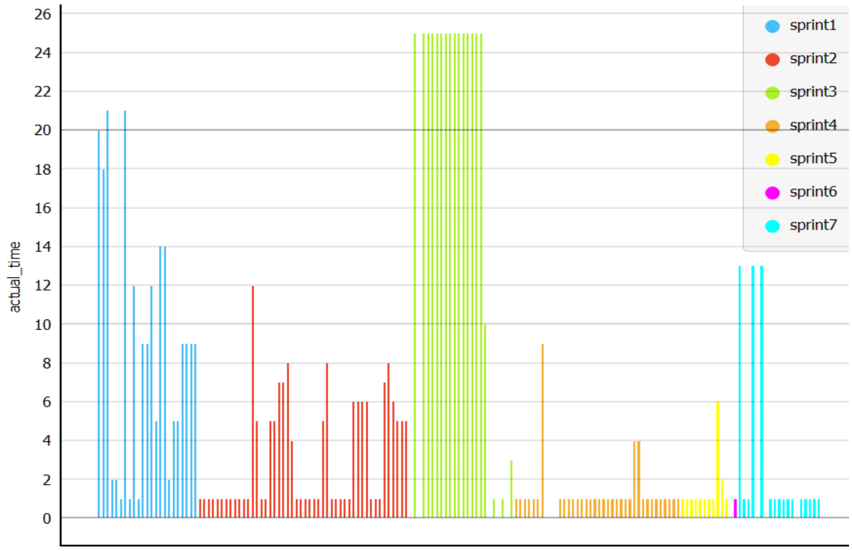

Figure 2. Bar plot of sprint completion time.

3.4. Modelling

The modeling stage is the stage of applying Machine Learning techniques of Decision Tree, Random Forest, Gradient Boosting, and AdaBoost algorithms with input from datasets that have been processed at the preprocessing stage, with split train-test data and k-fold validation. Model building is done in Orange Data Mining ver 3.32 by dividing into 6 scenarios, they are:

1. 70:30 data split without feature selection
2. 70:30 data split with Rrelieff feature selection
3. 70:30 data split with PCA feature selection
4. 10-fold validation without feature selection
5. 10-fold validation with Rrelieff feature selection
6. 10-fold validation with PCA feature selection

In the figure 3, the file is uploaded using a .csv file, preprocessing is done in the form of deleting rows with empty values, then sampling data 70:30, and applying the four Machine Learning models namely Decision Tree, Random Forest, Gradient Boosting, and AdaBoost, and then adding Rrelieff dan PCA feature selection separately.

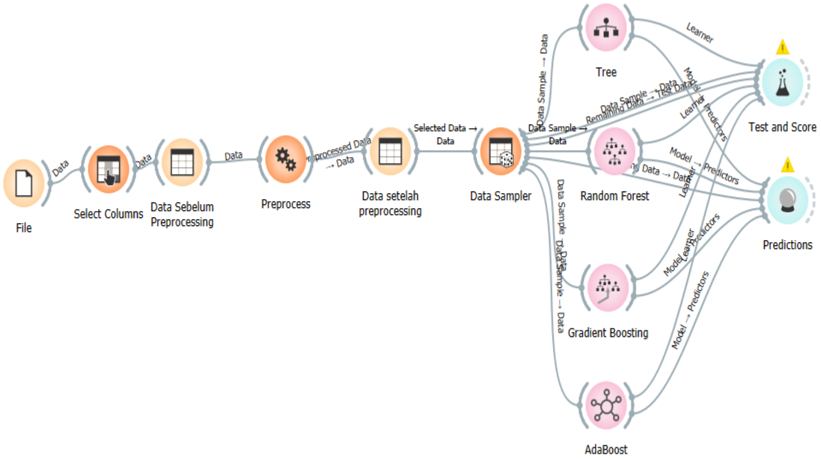

Figure 3. Modeling scenario with sampling data 70:30.

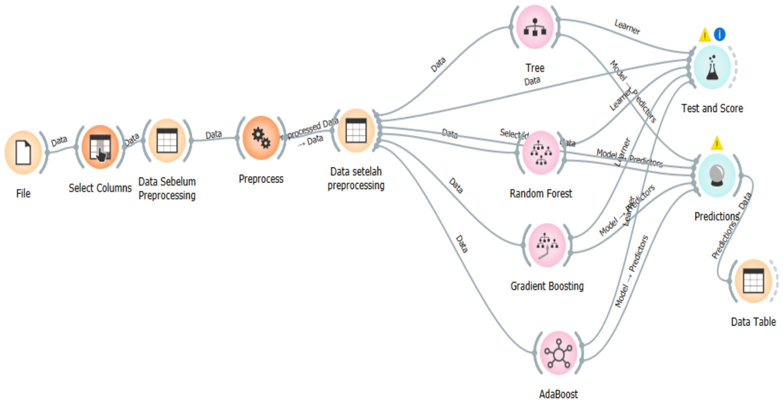

Figure 4. Modeling of scenario with 10-fold validation.

In the figure 4, the file is uploaded using a .csv file, preprocessing is done in the form of deleting rows with empty values, then applying 10-fold validation and the four Machine Learning models namely Decision Tree, Random Forest, Gradient Boosting, and AdaBoost, and then adding Rrelieff dan PCA feature selection separately.

3.5. Evaluation

Model performance is evaluated using MSE, RMSE, MAE, and R2 values. Based on the model results that have been tested and evaluated, the evaluation results are shown in the table 1 below:

Table 1. Machine Learning Algorithm Performance Evaluation Results

Model	**Split Data**	***feature selection***	**MSE**	**RMSE**	**MAE**	**R2**
Decision Tree	70:30	-	3.067	1.751	0.847	0.944
		Rrelieff	5.385	2.321	1.007	0.902
		PCA	3.142	1.773	0.839	0.943
	10-fold validation	-	3.622	1.903	0.979	0.938
		Rrelieff	4.307	2.075	1.020	0.926
		PCA	3.169	1.780	0.949	0.946
Random Forest	70:30	-	3.919	1.980	0.949	0.929
		Rrelieff	2.937	1.714	0.846	0.947
		PCA	3.178	1.783	0.846	0.942
	10-fold validation	-	3.244	1.801	0.931	0.945
		Rrelieff	3.440	1.855	0.934	0.941
		PCA	3.105	1.762	0.948	0.947
Gradient Boosting	70:30	-	3.169	1.780	0.866	0.942
		Rrelieff	3.084	1.756	0.852	0.944
		PCA	2.745	1.657	0.789	0.950
	10-fold validation	-	3.265	1.807	0.932	0.944
		Rrelieff	3.367	1.835	0.933	0.943
		PCA	2.895	1.701	0.898	0.951
AdaBoost	70:30	-	3.346	1.829	0.910	0.939
		Rrelieff	3.317	1.821	0.905	0.940
		PCA	3.416	1.848	0.878	0.938
	10-fold validation	-	4.353	2.086	1.123	0.926
		Rrelieff	4.986	2.233	1.150	0.915
		PCA	3.527	1.878	1.047	0.940

3.6. Deployment

Based on the evaluation results of 24 models, it is found that there is one best algorithm model, namely Gradient Boosting with 10-fold validation feature selection PCA. So that the model is used to perform deployment that can predict the time effort of software development with agile methods.

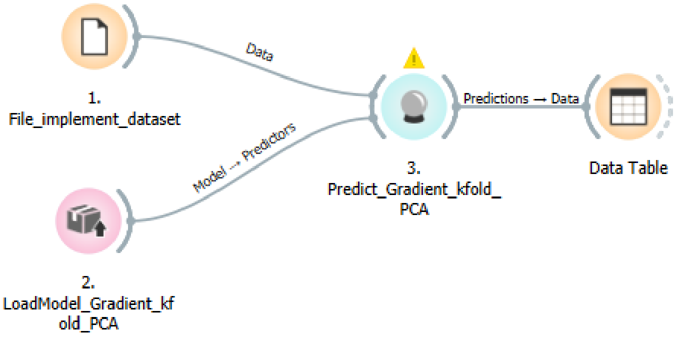

Figure 5. visualization of the best deployment model algorithm.

Based on the deployment in Figure 5, the prediction results are obtained with a sample of 5 rows as follows.

Table 2. Machine Learning Algorithm Performance Evaluation Results

No	Actual_time	Predicted_best_model
1	20	24.1089
2	18	24.1089
3	21	24.1089
4	2	5.24532
5	2	5.24532

The results of this deployment resulted in an MSE value of 1.968, RMSE of 1.403, MAE of 0.660, and R2 of 0.966.

4. RESULTS AND DISCUSSION

Based on table 1 above, the evaluation results can be visualized with the following figure 6 and 7.

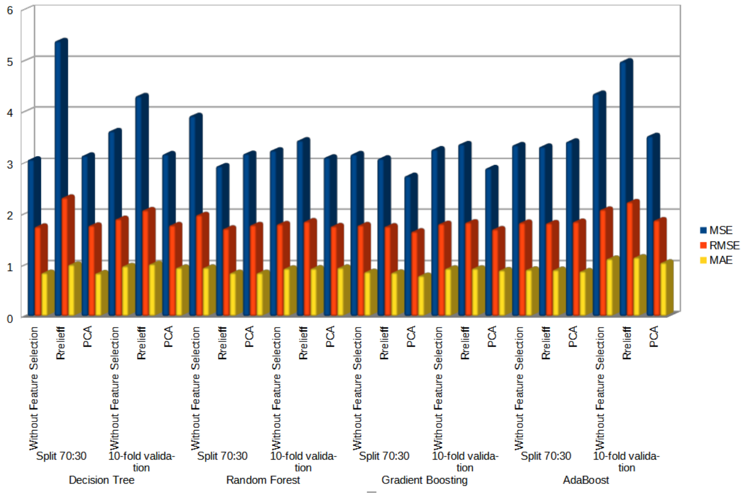

Figure 6. Performance graph of MSE, RMSE, and MAE of each model.

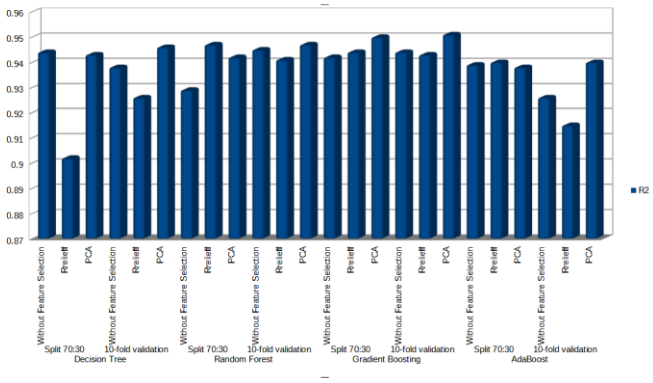

Figure 7. Performance graph of R2 of each model.

Based on the evaluation results of the 24 models above, it is found that the best machine learning algorithm performance used for this research dataset is the Gradient Boosting algorithm with 10-fold validation feature selection PCA, with performance values of MSE, RMSE, MAE, and R2 are 2.895, 1.701, 0.898, and 0.951. Figure 8 shows that the comparison between actual time effort and predicted time effort is linear.

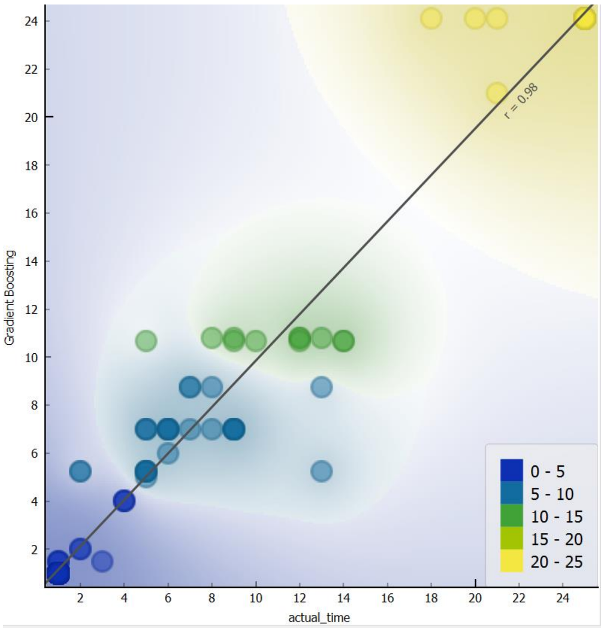

Figure 8. Comparison of actual and predicted time effort.

In this study, the use of Machine Learning techniques in calculating time effort on agile projects proved to be effective. The results show that the use of Feature Selection PCA can consistently improve the performance of almost all algorithm models in this study. The results of this study provide important insights in choosing the most suitable method for predicting time effort in software development using the agile approach.

5. CONCLUSIONS AND SUGGESTIONS

In this study, the use of Machine Learning techniques in calculating time effort on agile projects proved to be effective. Based on historical backlog data, Machine Learning is able to predict time effort quickly and accurately. This research tries four Machine Learning algorithms, namely Decision Tree, Random Forest, Gradient Boosting, and AdaBoost to predict time effort on agile projects. The results show that the use of Feature Selection PCA can consistently improve the performance of almost all algorithm models in this study. In this study, Gradient Boosting algorithm with 10-fold validation method and feature selection PCA, proved to have the best performance with MSE value of 2.895, RMSE 1.701, MAE 0.898, and R2 0.951. The results of this study provide important insights in choosing the most suitable method for predicting time effort in software development using the agile approach.

Some things that the author can suggest in the context of research development are:

1. Increase the number of datasets to produce better and more accurate tests and predictions.
2. Research can be developed using other Machine Learning algorithms with other feature selection optimizations.
3. Research can be developed by adding prediction features to the dataset.

REFERENCES

Baura, G. D. (2002). *Fuzzy Models*.

Castillo, F. (2016). *Managing information technology*. Springer.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

Hohl, P., Klünder, J., van Bennekum, A., Lockard, R., Gifford, J., Münch, J., … Schneider, K. (2018). Back to the future: origins and directions of the “Agile Manifesto”--views of the originators. *Journal of Software Engineering Research and Development*, *6*, 1–27.

Kementerian Keuangan RI. *Peraturan Menteri Keuangan Republik Indonesia Nomor 118/PMK.01/2021 tentang Organisasi dan Tata Kerja Kementerian Keuangan*. , (2021).

Khammas, B. M., Monemi, A., Bassi, J. S., Ismail, I., Nor, S. M., & Marsono, M. N. (2015). Feature selection and machine learning classification for malware detection. *Jurnal Teknologi*, *77*(1), 243–250.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249–256). Elsevier.

Mahnic, V., & Drnovscek, S. (2005). Agile software project management with scrum. *EUNIS 2005 Conference-Session Papers and Tutorial Abstracts*, 6.

Nassif, A. B., Capretz, L. F., & Ho, D. (2012). Estimating software effort using an ANN model based on use case points. *2012 11th International Conference on Machine Learning and Applications*, *2*, 42–47.

Peraturan Presiden Republik Indonesia. Tahun 2018 tentang Sistem Pemerintahan Berbasis Elektronik. , Peraturan Presiden Nomor § (2018).

Richards, J. D. (2007). *Book review: Introduction to IT project management*. SAGE Publications Sage CA: Los Angeles, CA.

Rodríguez Sánchez, E., Vázquez Santacruz, E. F., & Cervantes Maceda, H. (2023). Effort and Cost Estimation Using Decision Tree Techniques and Story Points in Agile Software Development. *Mathematics*, *11*(6), 1477.