ISSN: 2477-5126
e-ISSN: 2548-9356

Eksperimen Seleksi Fitur Pada Parameter Proyek Untuk *Software Effort Estimation* dengan K-*Nearest Neighbors*

Fachruddin[1], Yovi Pratama[2*)]
[1,2]Sekolah Tinggi llmu Komputer Dinamika Bangsa Jambi
[1,2]Jl. Jendral Sudirman, The Hok. Jambi Selatan. Jambi, Indonesia
email: [1]fachruddin_didin@yahoo.com, [2]yovi.pratama@gmail.com

***Abstract* – Software Effort Estimation is the process of estimating the cost of projects. Previous studies have been doing software estimation with some methods, both machine learning and non machine learning. This study using a set of project parameters with information gain and mutual information as feature selection and k-nearest neighbors to estimate project's effort. The dataset used in the experiment are albrecht, china, kemerer and mizayaki94 which can be obtained from a special data repository on the url http://openscience.us/repo/effort/. The authors develop the application to select project parameters to produces the arff dataset. This application is built in java language using IDE Netbean. Then the generated dataset is a selected parameter that will be compared to performe Software Effort Estimation using the WEKA tool. Feature Selection successfully decreases the estimated error value (represented by RAE and RMSE values). This means that the lower error value (RAE and RMSE) the more accurate the estimated cost. Estimation gets better after doing the selection of features either using information gain or mutual information. From the error value generated it can be concluded that the dataset generated using information gain method is better than mutual information but not too significant.**

***Abstrak –Software Effort Estimation* adalah proses estimasi biaya perangkat lunak sebagai suatu proses penting dalam melakukan proyek perangkat lunak. Berbagai penelitian terdahulu telah melakukan estimasi usaha perangkat lunak dengan berbagai metode, baik metode *machine learning* maupun non *machine learning*. Penelitian ini mengadakan set eksperimen seleksi atribut pada parameter proyek menggunakan teknik k-*nearest neighbours* sebagai estimasinya dengan melakukan seleksi atribut menggunakan *information gain* dan *mutual information* serta bagaimana menemukan parameter proyek yang paling representif pada *software effort estimation*. Dataset *software estimation effort* yang digunakan pada eksperimen adalah yakni albrecht, china, kemerer dan mizayaki94 yang dapat diperoleh dari repositori data khusus *Software Effort Estimation* melalui url http://openscience.us/repo/effort/. Selanjutnya peneliti melakukan pembangunan aplikasi seleksi atribut untuk menyeleksi parameter proyek. Sistem ini menghasilkan dataset arff yang telah diseleksi. Aplikasi ini dibangun dengan bahasa java menggunakan IDE Netbean. Kemudian dataset yang telah di-generate merupakan parameter hasil seleksi yang akan dibandingkan pada saat melakukan *Software Effort Estimation* menggunakan tool WEKA . Seleksi Fitur berhasil menurunkan nilai error estimasi (yang diwakilkan oleh nilai RAE dan RMSE). Artinya bahwa semakin rendah nilai error (RAE dan RMSE) maka semakin akurat nilai estimasi yang dihasilkan. Estimasi semakin baik setelah di lakukan seleksi fitur baik menggunakan *information gain* maupun *mutual information*. Dari nilai error yang dihasilkan maka dapat disimpulkan bahwa dataset yang dihasilkan seleksi fitur dengan metode *information gain* lebih baik dibanding *mutual information* namun, perbedaan keduanya tidak terlalu signifikan.**

Kata Kunci- ***Software Effort Estimation, k-nearest neighbors, information gain, mutual information*.**

*) penulis korespondensi (Yovi Pratama)
Email: yovi.pratama@gmail.com

I. PENDAHULUAN

Software Effort Estimation atau estimasi usaha perangkat lunak adalah bagian penting dari sebuah manajemen proyek. Estimasi yang akurat membantu kita menyelesaikan proyek dalam waktu dan anggaran yang telah ditentukan. Terdapat berbagai teknik, model estimasi dan tools untuk estimasi perangkat lunak [1]. Idealnya, dalam melakukan estimasi usaha perangkat lunak (sebagai salah satu rekayasa perangkat lunak) harus dapat menggunakan teknik *machine learning* untuk mengontrol atau secara signifikan mengurangi usaha terkait dengan membangun perangkat lunak [2][3].

Terdapat berbagai masalah dalam melakukan estimasi usaha pembangunan perangkat lunak. Salah satu masalah tersebut adalah kelangkaan contoh atau kurangnya data empiris dalam disiplin ilmu rekayasa perangkat lunak. Baik dalam estimasi usaha dengan pendekatan model ataupun dengan pendekatan machine learning. Sehingga ini menjadi suatu dilema dalam membangun model untuk memprediksi usaha suatu proyek perangkat lunak. Masalah lainnya adalah bahwa menemukan parameter proyek yang paling representative terhadap nilai effort yang dihasilkan. Beberapa penelitian telah menunjukkan bahwa tingkat akurasi estimasi usaha software sangat tergantung pada nilai-nilai parameter. Selain itu, telah ditunjukkan bahwa seleksi atribut memiliki pengaruh penting pada akurasi estimasi [4].

Terkait masalah tersebut, pada penelitian ini dilakukan pemilihan parameter proyek dari beberapa kumpulan data proyek yang tersedia dari berbagai sumber. Pemilihan parameter ini dilakukan untuk menemukan seberapa representatif nilai parameter pada proyek. Proses pemilihan ini sendiri dilakukan menggunakan metode seleksi atribut. *Information gain* dan *chi-square* adalah metode seleksi atribut yang paling efektif untuk meningkatkan akurasi pada beberapa algoritma machine learning. Seleksi atribut jenis *mutual information* merupakan metode yang paling umum yang digunakan dalam pemodelan bahasa statistik [5][6]. Sedangkan pada ekperimen komparasi antara seleksi atribut berbasis probabilitas dan berbasis frekuensi menghasilkan

bahwa seleksi atribut terbaik dihasilkan oleh gabungan *information gain* dan *mutual information* dengan nilai akurasi mencapai 91,36% [7].

Berdasarkan penelitian terdahulu mengenai seleksi atribut dan estimasi perangkat lunak, maka dalam penelitian ini mengadakan set eksperimen seleksi atribut pada parameter proyek menggunakan teknik *k-nearest neighbours* sebagai estimasinya. Sedangkan untuk metode seleksi atribut yang digunakan adalah *information gain* dan *mutual information*. Metode seleksi atribut yang digunakan dalam eksperimen ini akan menyeleksi dan mengkarakterisasi parameter atau atribut yang potensial, representatif, prediktif dalam sebuah estimasi usaha perangkat lunak. Hasil riset ini secara statistik dinilai dan dievaluasi dengan model *evaluasi mean absolute error* dan *root mean absolute error*.

II. PENELITIAN YANG TERKAIT

Berbagai penelitian terdahulu telah melakukan estimasi usaha perangkat lunak dengan berbagai metode, baik metode *machine learning* maupun non *machine learning*. Beberapa metode estimasi usaha perangkat lunak dengan metode non *machine learning* antara lain SLIM, COCOMO, function point, use case point, dan metode analogi sherpped. Penelitian terkait estimasi usaha perangkat lunak yang dikembangkan dengan menggunakan berbagai *machine learning* antara lain penggunaan jaringan syaraf tiruan[7][8], *support vector machine* [7][14][15][16], *naïve bayes* [4][5], k-*nearest neighbor* [7][14][15][16], *data mining* dan *fuzzy*, *case based reasoning* [11] serta *liner regression* [14][15].

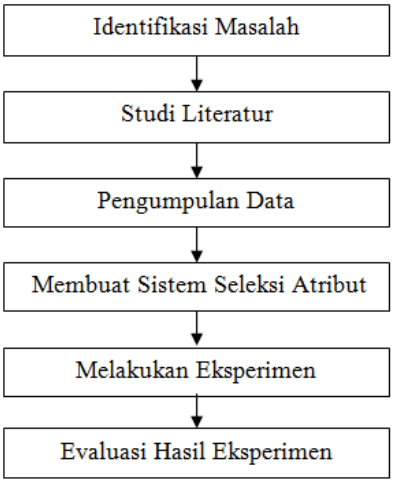
Penelitian sebelumnya yang telah dilakukan untuk estimasi usaha perangkat lunak dengan *machine learning* menyatakan bahwa penggunaan teknik *machine learning* lebih baik dari pada non machine learning. Hasil estimasi usaha perangkat lunak terbaik adalah dengan menggunakan kNN sebagai estimatornya dengan nilai RMSE sebesar 7.5 [7] Penelitian tersebut melakukan komparasi beberapa metode *machine learning* dengan teknik *neural network*, *k- nearest neighbor* & *support vector machine*. Untuk metode non *machine learning* yang digunakan penelitian tersebut adalah *use case point* dan *function point*. Penelitian lainnya menyatakan bahwa metode non *machine learning* yakni pendekatan analogi mampu melakukan estimasi usaha dan biaya dengan lebih baik. Teknik analogi yang telah dimodifikasi mampu menghasilkan nilai estimasi yang lebih baik dibanding analogi sherpped dengan *residual error* (RE) rata-rata sebesar 17 % pada 10 jenis proyek dari *dataset* [1].

III. METODE PENELITIAN

Pada penelitian ini peneliti menggunakan tahapan kegiatan penelitian sebagai berikut :

A. *Kerangka Kerja Penelitian*

Pada penelitian ini peneliti menggunakan tahapan kerangka kerja penelitian sebagai berikut :



Gbr. 1 Kerangka Kerja Penelitian

1) Identifikasi Masalah

Pada tahap ini peneliti melakukan observasi awal untuk mengidentifikasi dan merumuskan masalah dari berbagai penelitian yang telah dilakukan terkait software effort estimation. Masalah yang dibahas adalah bagaimana memilih parameter proyek dengan melakukan seleksi atribut menggunakan *information gain* dan *mutual information* serta bagaimana menemukan parameter proyek yang paling representif pada *Software Effort Estimation* dengan menggunakan algoritma *k-nearest neighbor*.

2) Studi literatur

Mempelajari dan memahami berbagai teori tentang *software estimation effort*, *information gain*, *mutual information* dan k-*nearest neighbor* yang akan diujikan guna melakukan eksperimen yang dibahas dalam penelitian ini.

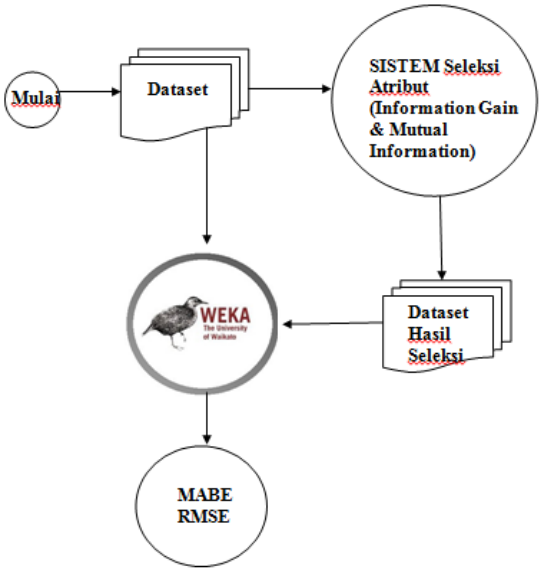
3) Pengumpulan Data

Kegiatan yang dilakukan pada tahap ini adalah mengumpulkan *dataset yang digunakan. Dataset software estimation effort* yang digunakan pada eksperimen adalah yakni albrecht, china, kemerer dan mizayaki94 yang dapat diperoleh dari repositori data khusus *Software Effort Estimation* melalui url http://openscience.us/repo/effort/. Kemudian melakukan pengamatan pada dataset tersebut.

4) Membuat Sistem Seleksi Atribut

Kegiatan ini berupa pembangunan aplikasi seleksi atribut untuk menyeleksi parameter proyek. Sistem ini menghasilkan dataset arff yang telah diseleksi. Aplikasi ini dibangun dengan bahasa java menggunakan IDE *Netbean*. Kemudian dataset yang telah di-generate merupakan parameter hasil seleksi yang akan dibandingkan pada saat melakukan *software effort estimation* :

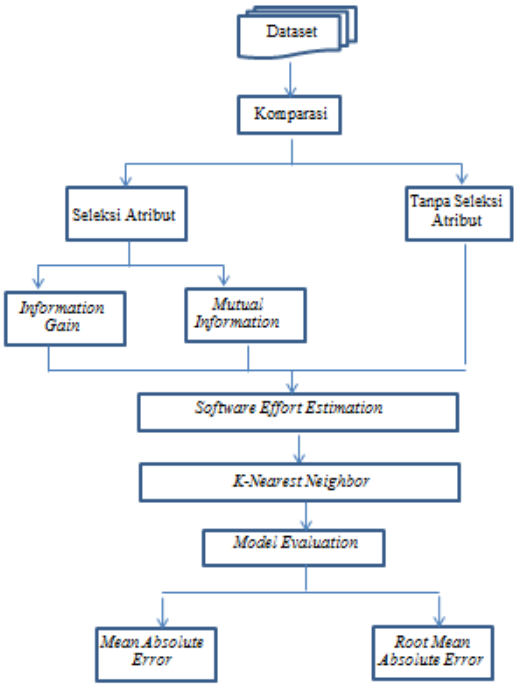
ISSN: 2477-5126
e-ISSN: 2548-9356



Gbr. 2 Skema Sistem Eksperimen

5) Melakukan eksperimen

Kegiatan ini melakukan eksperimen pada dataset yang telah dihasilkan dengan sistem seleksi atribut dengan *information gain* dan *mutual information* kemudian dibandingkan dengan dataset yang belum diseleksi. Gambar 3.3 berikut ini adalah alur skema eskperimen yang dilakukan dalam penelitian ini:



Gbr. 3Alur Skenario Eksperimen

6) Evaluasi hasil Eksperimen

Kegiatan ini adalah melakukan evaluasi dan analisis terhadap hasil estimasi proyek perangkat lunak dengan menggunakan tools *machine learning* WEKA. Hasil evaluasi yang dianalisis adalah *mean absolute error* (RME) dan *root mean squared error* (RMSE).

B. Bahan Penelitian

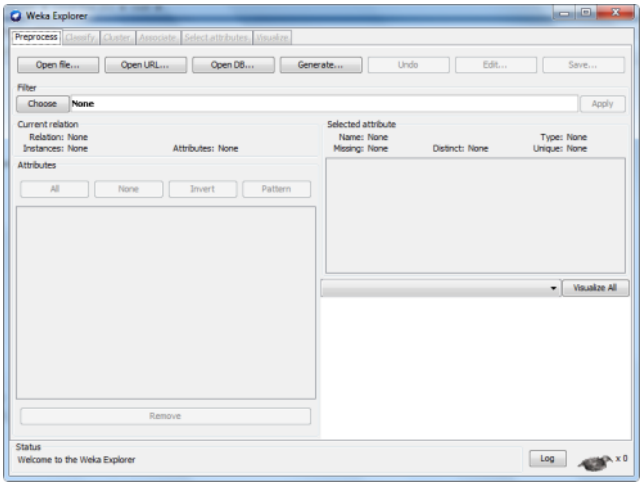
Bahan penelitian yang dibutuhkan yaitu:

1. Alur Eksperimen dan scenario yang telah dipersiapkan.
2. Dataset Albrecth, China, Mizayaki94 dan Kemerer.
3. Algoritma Seleksi Fitur *Information gain* dan Mutual Information.
4. Algoritma K-*Nearest Neighbor.*

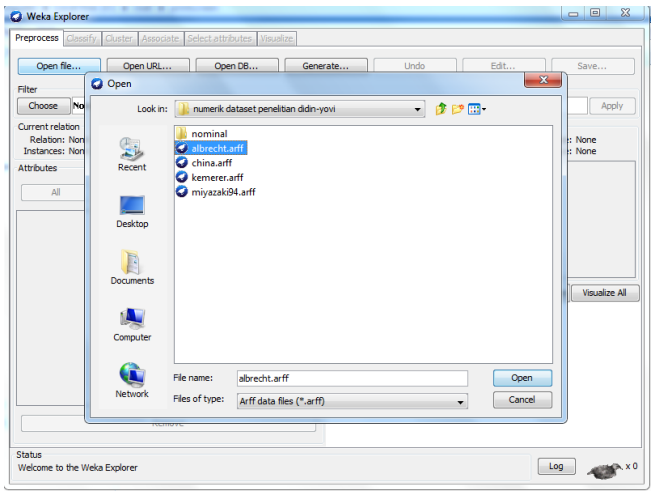
IV. HASIL DAN PEMBAHASAN

A. Analisa Hasil Konversi Dataset Numerik Menjadi Dataset Nominal

Pada tahap ini perlu dilakuakn konversi dataset numerik menjadi dataset nominal, karena dalam melakukan estimasi biaya software menggunakan K-*nearest Neighbor* harus menggunakan dataset tipe nominal. WEKA mampu menghandle tugas tersebut. Proses perubahan dataset numerik menjadi dataset nominal tersebut dilakukan secara otomatis oleh WEKA tools. Gambar 4,5,6,7 dan 8 berikut ini adalah langkah konversi data dari numerik menjadi nominal.

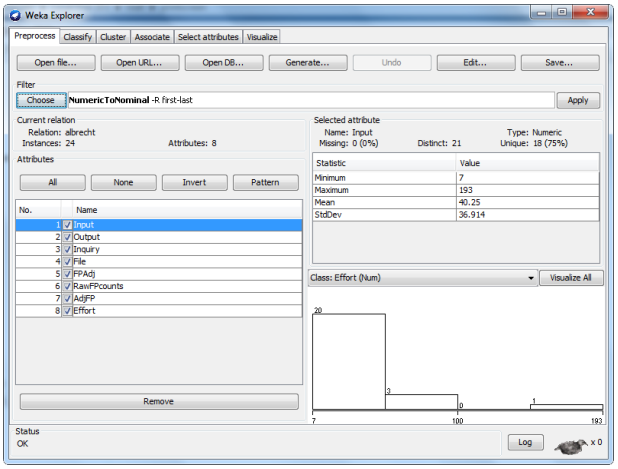


Gbr. 4 *Interface* Input data untuk melakukan Konversi dataset

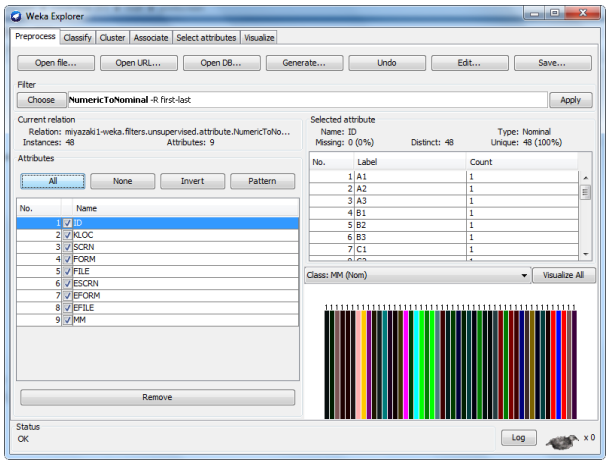


Gbr. 5 Interface Ambil Data dari Direktori

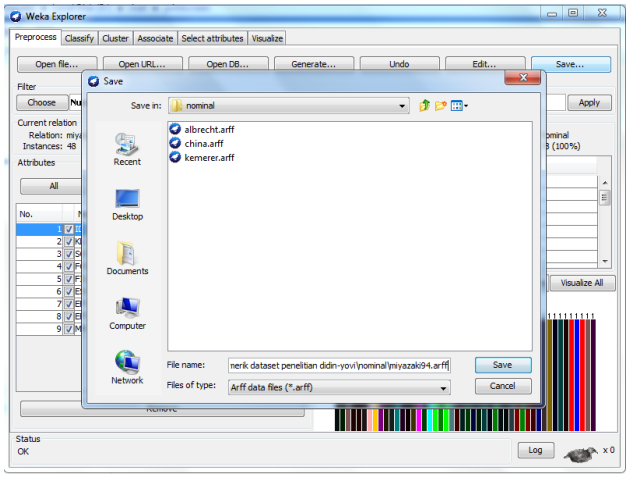
ISSN: 2477-5126
e-ISSN: 2548-9356



Gbr. 6 Interface Konversi dataset Numerik menjadi dataset Nominal



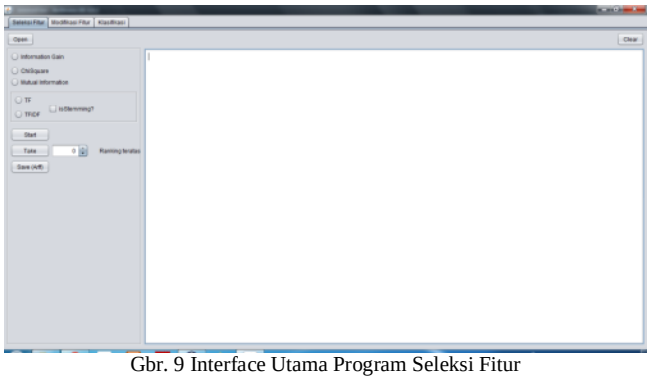
Gbr. 7 Visualisasi Dataset Nominal



Gbr. 8 Penyimpanan Dataset Nominal

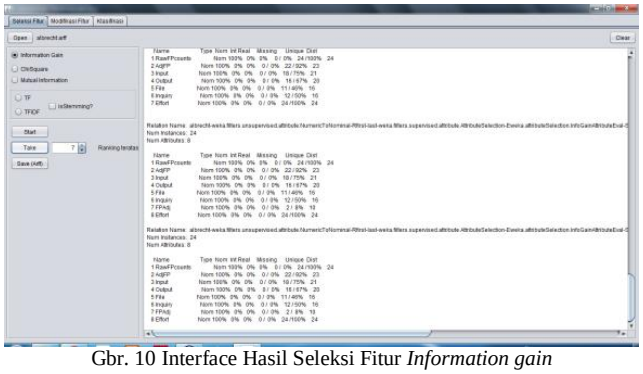
B. *Analisa Hasil Dataset Dengan Seleksi Fitur*

Setelah dilakukan konversi *dataset* dari tipe numerik menjadi nominal, maka tahap selanjutnya adalah melakukan seleksi fitur menggunakan algoritma *information gain* dan *mutual information.* Selanjutnya untuk melakukan tugas tersebut. Maka dalam penelitian ini dilakukan pembangunan program khusus untuk melakukan task Seleksi fitur dengan algoritma *information gain* dan *mutual information.* Program yang digunakan dibangun dengan Bahasa JAVA mennggunakan IDE Netbean 8.0.



Gbr. 9 Interface Utama Program Seleksi Fitur

C. *Analisa Hasil Seleksi Fitur Dengan Information Gain*



Gbr. 10 Interface Hasil Seleksi Fitur *Information gain*

D. *Analisa Hasil Information gain Pada Dataset Albrecth*

TABEL I
HASIL INFORMATION GAIN PADA DATASET ALBRECTH

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai Information Gain	Status
1	Input	RawFPcounts	4.585	Parameter Proyek/ Atribut Proyek
2	Output	AdjFP	4.502	
3	Inquiry	Input	4.335	
4	File	Output	4.252	
5	FPAdj	File	3.768	
6	RawFPcounts	Inquiry	3.736	
7	AdjFP	FPAdj	3.209	
8	**Effort**	**Effort**		Estimasi

Tabel 1 di atas adalah hasil seleksi fitur menggunakan *Information Gain* pada dataset Albretch. Hasil seleksi fitur tersebut menghasilkan urutan parameter yang berbeda dengan sebelum diseleksi.

E. Analisa Hasil Information gain Pada Dataset China

TABEL II
HASIL INFORMATION GAIN PADA DATASET CHINA

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai Information Gain	Status
1	ID	ID	8.899	Parameter Proyek/ Atribut Proyek
2	AFP	N_effort	8.857	
3	Input	AFP	8.2	
4	Output	Added	7.756	
5	Enquiry	NPDR_AFP	7.542	
6	File	NPDU_UFP	7.532	
7	Interface	PDR_AFP	7.498	
8	Added	PDR_UFP	7.469	
9	Changed	Input	7.29	
10	Deleted	Output	6.804	
11	PDR_AFP	Enquiry	6.16	
12	PDR_UFP	File	6.01	
13	NPDR_AFP	Changed	4.609	
14	NPDU_UFP	Duration	4.275	
15	Resource	Interface	3.368	
16	Dev.Type	Resource	1.198	
17	Duration	Deleted	1.193	
18	**Effort**	**Effort**	-	**Estimasi**

Tabel 2 di atas adalah hasil seleksi fitur menggunakan *Information Gain* pada dataset China. Hasil seleksi fitur tersebut menghasilkan urutan parameter yang berbeda dengan sebelum diseleksi .

F. Analisa Hasil Information gain Pada Dataset Mizayaki94

TABEL III
HASIL *INFORMATION GAIN* PADA DATASET MIZAYAKI94

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai Information Gain	Status
1	ID	ID	5.418	Parameter Proyek/ Atribut Proyek
2	KLOC	EFILE	5.377	
3	SCRN	EFORM	5.293	
4	FORM	KLOC	5.293	
5	FILE	ESCRN	5.21	
6	ESCRN	SCRN	4.861	
7	EFORM	FORM	4.762	
8	EFILE	FILE	4.736	
9	**MM**	**MM**	-	**Estimasi**

Tabel 3 di atas adalah hasil seleksi fitur menggunakan *Information Gain* pada dataset Mizayaki94. Hasil seleksi fitur tersebut menghasilkan urutan parameter yang berbeda dengan sebelum diseleksi.

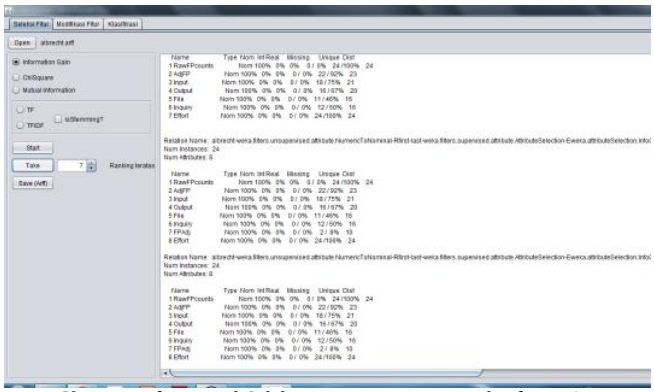
G. Analisa Hasil Information gain Pada Dataset Kemerer

TABEL IV
HASIL *INFORMATION GAIN* PADA DATASET KEMERER

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai Information Gain	Status
1	ID	RAWFP	3.907	Parameter Proyek/ Atribut Proyek
2	Language	AdjFP	3.907	
3	Hardware	ID	3.907	
4	Duration	KSLOC	3.907	
5	KSLOC	Duration	3.323	
6	AdjFP	Hardware	2.146	
7	RAWFP	Language	0.7	
8	**EffortMM**	**EffortMM**	-	Estimasi

Tabel 4 di atas adalah hasil seleksi fitur menggunakan *Information Gain* pada dataset Albretch. Hasil seleksi fitur tersebut menghasilkan urutan parameter yang berbeda dengan sebelum diseleksi.

H. Analisa Hasil Seleksi Fitur Dengan Mutual Information



Gbr 1 *Interface* Hasil Seleksi Fitur Dengan *Mutual Information*

I. Analisa Hasil Mutual information Pada Dataset Albrecth

TABEL V
HASIL *MUTUAL INFORMATION PADA DATASET ALBRECTH*

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai Mutual Information	Status
1	Input	RawFPcounts	4.585	Parameter Proyek/ Atribut Proyek
2	Output	AdjFP	4.502	
3	Inquiry	Input	4.335	
4	File	Output	4.252	
5	FPAdj	File	3.768	
6	RawFPcounts	Inquiry	3.736	
7	AdjFP	FPAdj	3.209	
8	**Effort**	**Effort**	-	**Estimasi**

Tabel 5 di atas adalah hasil seleksi fitur menggunakan *Mutual information* pada dataset Albretch. Hasil seleksi fitur tersebut menghasilkan urutan parameter yang berbeda dengan sebelum diseleksi.

J. Analisa Hasil Mutual information Pada Dataset China

TABEL VI
HASIL MUTUAL INFORMATION PADA DATASET CHINA

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai MI	Status
1	ID	ID	8.899	Parameter Proyek/ Atribut Proyek
2	AFP	N_effort	8.857	
3	Input	AFP	8.2	
4	Output	Added	7.756	
5	Enquiry	NPDR_AFP	7.542	
6	File	NPDU_UFP	7.532	
7	Interface	PDR_AFP	7.498	
8	Added	PDR_UFP	7.469	
9	Changed	Input	7.29	
10	Deleted	Output	6.804	
11	PDR_AFP	Enquiry	6.16	
12	PDR_UFP	File	6.01	
13	NPDR_AFP	Changed	4.609	
14	NPDU_UFP	Duration	4.275	
15	Resource	Interface	3.368	
16	Dev.Type	Resource	1.198	
17	Duration	Deleted	1.193	
18	N_effort	Dev.Type	0	
19	**Effort**	**Effort**	-	**Estimasi**

ISSN: 2477-5126
e-ISSN: 2548-9356

Tabel 6 di atas adalah hasil seleksi fitur menggunakan *Mutual information* pada dataset China. Hasil seleksi fitur tersebut menghasilkan urutan parameter yang berbeda dengan sebelum diseleksi.

K. Analisa Mutual information Pada DatasetMizayaki94

TABEL VII
HASIL *MUTUAL INFORMATION* PADA DATASET MIZAYAKI94

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai Mutual Information	Status
1	ID	ID	5.418	Parameter Proyek/ Atribut Proyek
2	KLOC	EFILE	5.377	
3	SCRN	KLOC	5.293	
4	FORM	EFORM	5.293	
5	FILE	ESCRN	5.21	
6	ESCRN	SCRN	4.861	
7	EFORM	FORM	4.762	
8	EFILE	FILE	4.736	
9	**MM**	**MM**	-	**Estimasi**

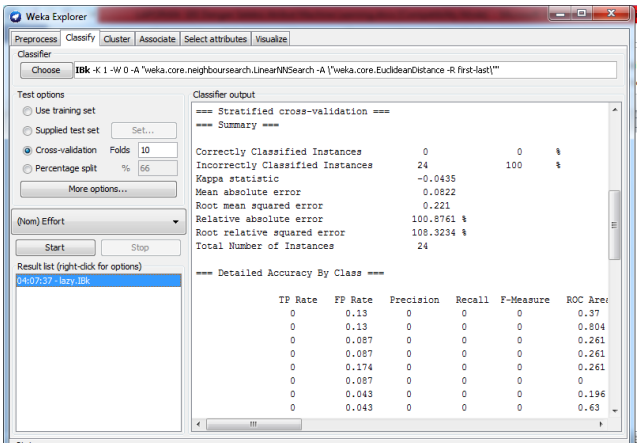
Tabel 7 di atas adalah hasil seleksi fitur menggunakan *Mutual information* pada dataset Mizayaki94. Hasil seleksi fitur tersebut menghasilkan urutan parameter yang berbeda dengan sebelum diseleksi

L. Analisa Hasil Mutual information Pada Dataset Kemerer

TABEL VIII
HASIL *MUTUAL INFORMATION* PADA DATASET KEMERER

NO	Rangking Parameter Sebelum di Seleksi	Rangking Parameter Setelah di Seleksi	Nilai Mutual Information	Status
1	ID	ID	3.907	Parameter Proyek/ Atribut Proyek
2	Language	RAWFP	3.907	
3	Hardware	AdjFP	3.907	
4	Duration	KSLOC	3.907	
5	KSLOC	Duration	3.323	
6	AdjFP	Hardware	2.146	
7	RAWFP	Language	0.7	
8	**EffortMM**	**EffortMM**	-	**Estimasi**

M. Analisa Hasil Estimasi DataSet



Gbr. 10 Interface Hasil Estimasi DataSet

Tabel 8 di atas adalah hasil estimasi dari seluruh dataset yang digunakan. Hasil estimasi berupa nilai evaluasi error antara lain RAE dan RMSE. Untuk hasil RAE dan RMSE detail setiap kombinasi dataset dengan jenis seleksi fitur dapat dilihat pada sub bab berikut.

N. Perbandingan Analisa Hasil Estimasi DataSet Dengan Information Gain

TABEL IX
HASIL ESTIMASI (DENGAN *INFORMATION GAIN*) DATASET ALBRETCH

NO	Jumlah Parameter	Parameter	MAE	RMSE	Estimasi
1	Dataset Albrecth (Tanpa Seleksi Fitur)	7	6.1917	9.5831	**MAE Terkecil** =0.0822 **RMSE Terkecil=** 0.2084
2	Dataset Albrecth (Dengan Seleksi Fitur Information Gain)	7	**0.0822**	0.221	
		6	**0.0822**	0.2201	
		5	0.0823	0.2183	
		4	0.0824	0.2165	
		3	0.0827	0.2132	
		2	.083	**0.2098**	
		1	0.0832	0.2084	

Dari tabel 9 tersebut dapat dilihat nilai MAE dan RMSE dari dataset Albrecth hasil pemilihan fitur menggunakan *Infromation Gain.* Nilai MAE terkecil adalah 0.0822 dan RMSE terkecil adalah 0.2084.

TABEL X
HASIL ESTIMASI (DENGAN *INFORMATION GAIN*) DATASET CHINA

NO	Jumlah Parameter	Parameter	MAE	RMSE	Estimasi
1	(Tanpa Seleksi Fitur)	18	15.1824	30.7818	**MAE Terkecil** =0.0041 **RMSE Terkecil** = 0.0452
2	(Dengan Seleksi Fitur Information Gain)	18	**0.0041**	0.0495	
		15	**0.0041**	0.0494	
		13	**0.0041**	0.0493	
		10	**0.0041**	0.0492	
		9	**0.0041**	0.0493	
		8	**0.0041**	0.0494	
		7	**0.0041**	0.0493	
		6	**0.0041**	0.0487	
		5	**0.0041**	0.0485	
		4	**0.0041**	0.0479	
		3	**0.0041**	0.0473	
		2	**0.0041**	**0.0452**	
		1	**0.0041**	0.0455	

TABEL XI
HASIL ESTIMASI (DENGAN *INFORMATION GAIN*) DATASET MIZAYAKI94

NO	Jumlah Parameter	Parameter	MAE	RMSE	Estimasi
1	(Tanpa Seleksi Fitur)	8	59.5729	20.635	**MAE Terkecil** =0.0451 **RMSE Terkecil** = 0.1521
2	(Dengan Seleksi Fitur Information Gain)	8	0.0451	0.1647	
		7	0.0452	0.1622	
		6	0.0452	0.1601	
		5	0.0452	0.1568	
		4	0.0452	0.1557	
		3	0.0452	0.1541	
		2	0.0452	0.1528	
		1	0.0453	0.1521	

Dari tabel 11 tersebut dapat dilihat nilai MAE dan RMSE dari dataset Mizayaki94 hasil pemilihan fitur menggunakan *Infromation Gain.* Nilai MAE terkecil adalah 0. 0451 dan RMSE terkecil adalah 0. 1521.

TABEL XII
HASIL ESTIMASI (DENGAN *INFORMATION GAIN*) DATASET KEMERER

NO	Jumlah Parameter	Parameter	MAE	RMSE	Estimasi
1	(Tanpa Seleksi Fitur)	7	14.054	22.6402	**MAE Terkecil** =0.1307 **RMSE Terkecil=** 0.2663
2	(Dengan Seleksi Fitur Information Gain)	7	0.1309	0.2759	
		6	**0.1307**	0.276	
		5	0.1309	0.2743	
		4	0.1326	**0.2663**	
		3	0.1326	**0.2663**	
		2	0.1326	**0.2663**	
		1	0.1326	**0.2663**	

Dari tabel 12 tersebut dapat dilihat nilai MAE dan RMSE dari dataset Kemerer hasil pemilihan fitur menggunakan *Infromation Gain.* Nilai MAE terkecil adalah 0. 1307dan RMSE terkecil adalah 0. 2663.

O. Perbandingan Analisa Hasil Estimasi DataSet Dengan Mutual information

TABEL XIII
HASIL ESTIMASI (DENGAN *MUTUAL INFORMATION*)DATASET ALBRETCH

NO	Jumlah Parameter	Paret er	MAE	RMSE	Estimasi
1	(Tanpa Seleksi Fitur)	187	6.1917	9.5831	**MAE Terkecil** =0.0822 **RMSE Terkecil=** 0.2084
2	(Dengan Seleksi Fitur Mutual Information)	7	**0.0822**	0.221	
		6	**0.0822**	0.2201	
		5	0.0823	0.2183	
		4	0.0824	0.2165	
		3	0.0827	0.2132	
		2	0.083	0.2098	
		1	0.0832	**0.2084**	

Dari tabel 13 tersebut dapat dilihat nilai MAE dan RMSE dari dataset Albretch hasil pemilihan fitur menggunakan *Mutual Information.* Nilai MAE terkecil adalah 0. 0822 dan RMSE terkecil adalah 0. 2084.

TABEL XIV
HASIL ESTIMASI (DENGAN *MUTUAL INFORMATION*) DATASET CHINA

NO	Jumlah Parameter	Parameter	MAE	RMSE	Estimasi
1	(Tanpa Seleksi Fitur)	18	15.1824	30.7818	**MAE Terkecil** =0.0041 **RMSE Terkecil=** 0.0452
2	(Dengan Seleksi Fitur Mutual Information)	18	**0.0041**	0.0495	
		15	**0.0041**	0.0494	
		13	**0.0041**	0.0493	
		10	**0.0041**	0.0492	
		9	**0.0041**	0.0493	
		8	**0.0041**	0.0494	
		7	**0.0041**	0.0493	
		6	**0.0041**	0.0487	
		5	**0.0041**	0.0485	
		4	**0.0041**	0.0479	
		3	**0.0041**	0.0473	
		2	**0.0041**	**0.0452**	
		1	**0.0041**	0.0455	

Dari tabel 14 tersebut dapat dilihat nilai MAE dan RMSE dari dataset China hasil pemilihan fitur menggunakan *Mutual Information.* Nilai MAE terkecil adalah 0. 0041 dan RMSE terkecil adalah 0. 0452.

TABEL XV
HASIL ESTIMASI (DENGAN *MUTUAL INFORMATION*) DATASET MIZAYAKI94

NO	Jumlah Parameter	Parameter	MAE	RMSE	Estimasi
1	(Tanpa Seleksi Fitur)	8	59.5729	20.635	**MAE Terkecil** =0.0451 **RMSE Terkecil** = 0.1521
2	(Dengan Seleksi Fitur Mutual Information)	8	**0.0451**	0.1647	
		7	0.0452	0.1622	
		6	0.0452	0.1601	
		5	0.0452	0.1568	
		4	0.0452	0.1557	
		3	0.0452	0.1547	
		2	0.0452	0.1528	
		1	0.0453	**0.1521**	

Dari tabel 15 tersebut dapat dilihat nilai MAE dan RMSE dari dataset China hasil pemilihan fitur menggunakan *Mutual Information.* Nilai MAE terkecil adalah 0. 0451 dan RMSE terkecil adalah 0. 1521.

TABEL XVI
HASIL ESTIMASI (DENGAN *MUTUAL INFORMATION*) DATASET KEMERER

NO	Jumlah Parameter	Parameter	MAE	RMSE	Estimasi
1	(Tanpa Seleksi Fitur)	7	14.054	2 .6402	**MAE Terkecil** =0.1307 **RMSE Terkecil** = 0.2663
2	(Dengan Seleksi Fitur Mutual Information)	7	0.1309	0.2759	
		6	**0.1307**	0.276	
		5	0.1309	0.2743	
		4	0.1326	**0.2663**	
		3	0.1326	**0.2663**	
		2	0.1326	**0.2663**	
		1	0.1326	**0.2663**	

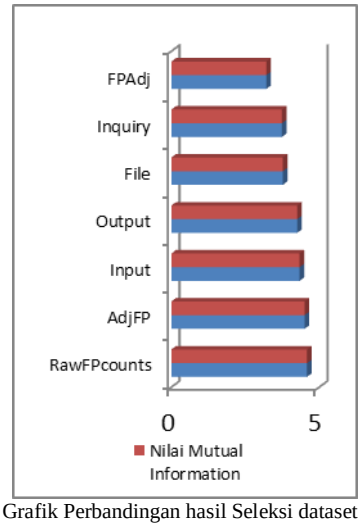
ISSN: 2477-5126
e-ISSN: 2548-9356

Dari tabel 16 tersebut dapat dilihat nilai MAE dan RMSE dari dataset Kemerer hasil pemilihan fitur menggunakan *Mutual Information*. Nilai MAE terkecil adalah 0. 1307dan RMSE terkecil adalah 0. 2663.

P. Perbandingan hasil Seleksi Information gain dan Mutual Information

TABEL XVII
HASIL PERBANDINGAN HASIL SELEKSI DATASET ALBRECTH

NO	Rangking Parameter	Nilai Information Gain	Nilai Mutual Information
1	RawFPcounts	4.585	4.585
2	AdjFP	4.502	4.502
3	Input	4.335	4.335
4	Output	4.252	4.252
5	File	3.768	3.768
6	Inquiry	3.736	3.736
7	FPAdj	3.209	3.209

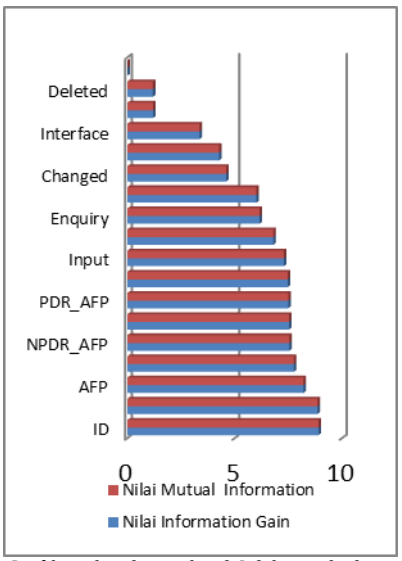


Gbr. 11 Grafik Perbandingan hasil Seleksi dataset Albrecth

Dari tabel 17 dan grafik pada gambar 11 dapat dilihat hasil perbandingan dataset Albrecth dengan seleksi fitur *information gain* dan *mutual information*. Dari hasil perbandingan tersebut tidak terdapat perbedaan rangking parameter hasil dua jenis seleksi fitur.

TABEL XVIII
HASIL PERBANDINGAN HASIL SELEKSI DATASET CHINA

No	Rangking Parameter	Nilai Information Gain	Nilai Mutual Information
1	ID	8.899	8.899
2	N_effort	8.857	8.857
3	AFP	8.2	8.2
4	Added	7.756	7.756
5	NPDR_AFP	7.542	7.542
6	NPDU_UFP	7.532	7.532
7	PDR_AFP	7.498	7.498
8	PDR_UFP	7.469	7.469
9	Input	7.29	7.29
10	Output	6.804	6.804
11	Enquiry	6.16	6.16
12	File	6.01	6.01
13	Changed	4.609	4.609
14	Duration	4.275	4.275
15	Interface	3.368	3.368
16	Resource	1.198	1.198
17	Deleted	1.193	1.193
18	Dev.Type	0	0

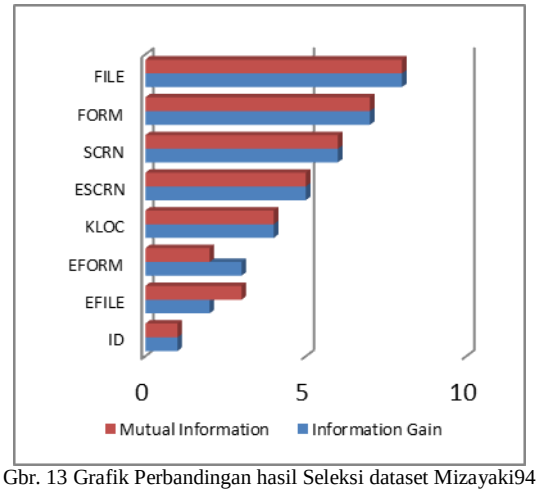


Gbr. 12 Grafik Perbandingan hasil Seleksi Pada dataset China

Dari tabel 18 dan grafik pada gambar 12 dapat dilihat hasil perbandingan dataset China dengan seleksi fitur *information gain* dan *mutual information*. Dari hasil perbandingan tersebut tidak terdapat perbedaan rangking parameter hasil dua jenis seleksi fitur.

TABEL XIX
HASIL PERBANDINGAN HASIL SELEKSI DATASET MIZAYAKI94

NO	Information Gain		Mutual Information	
	Parameter	Nilai	Parameter	Nilai
1	ID	5.418	ID	5.418
2	EFILE	5.377	EFILE	5.377
3	EFORM	5.293	KLOC	5.293
4	KLOC	5.293	EFORM	5.293
5	ESCRN	5.21	ESCRN	5.21
6	SCRN	4.861	SCRN	4.861
7	FORM	4.762	FORM	4.762
8	FILE	4.736	FILE	4.736



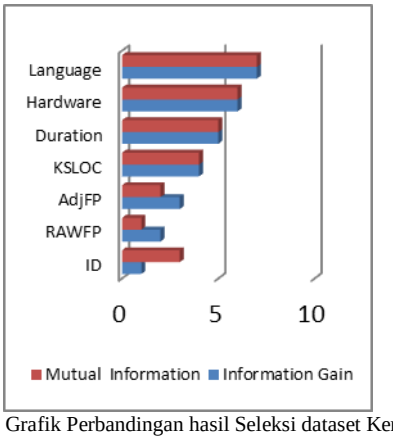
Gbr. 13 Grafik Perbandingan hasil Seleksi dataset Mizayaki94

Dari tabel 19 dan grafik pada gambar 13 dapat dilihat hasil perbandingan dataset China dengan seleksi fitur *information gain* dan *mutual information*. Dari hasil perbandingan tersebut perbedaan rangking parameter hasil

dua jenis seleksi fitur. Yakni pada Parameter EFFORM dan EFILE.

TABEL XX
HASIL PERBANDINGAN HASIL SELEKSI DATASET KEMERER

	Information Gain		Mutual Information	
NO	**Parameter**	**Nilai**	**Parameter**	**Nilai**
1	ID	3.907	RAWFP	3.907
2	RAWFP	3.907	AdjFP	3.907
3	AdjFP	3.907	ID	3.907
4	KSLOC	3.907	KSLOC	3.907
5	Duration	3.323	Duration	3.323
6	Hardware	2.146	Hardware	2.146
7	Language	0.7	Language	0.7



Gbr. 14 Grafik Perbandingan hasil Seleksi dataset Kemerer

Dari tabel 20 dan grafik pada gambar 14 dapat dilihat hasil perbandingan dataset China dengan seleksi fitur *information gain* dan *mutual information*. Dari hasil perbandingan tersebut perbedaan rangking parameter hasil dua jenis seleksi fitur. Yakni pada Parameter EFFORM dan EFILE.

Q. Analisis Keseluruhan Dataset

TABEL XXI
HASIL PERBANDINGAN ANALISIS KESELURUHAN DATASET

NO	**Dataset**	**Seleksi Fitur**	**Tipe ***	**RAE**	**RMSE**
1	Albrecth (7 parameter)	Tanpa	Nu	6.1917	9.5831
		Information Gain	No	0.0822	0.221
		Mutual Information	No	0.0822	0.221
2	China(18 parameter)	Tanpa	Nu	15.1824	30.7818
		Information Gain	No	0.0041	0.0495
		Mutual Information	No	0.0041	0.0495
3	Mizayaki94 (8 parameter)	Tanpa	Nu	59.5729	20.635
		Information Gain	No	0.0451	0.1647
		Mutual Information	No	0.0451	0.1647
4	Kemerer (7 parameter)	Tanpa	Nu	14.054	22.6402
		Information Gain	No	0.1309	0.2759
		Mutual Information	No	0.1309	0.2759

*Nu: Numerik ; No: Nominal

Pada tabel 21 di atas terdapat 2 jenis dataset yang ditampilkan dari dataset yang sama. Tipe tersebut adalah tipe data nominal dan numerik. Tipe nominal merupakan hasil konversi dari dataset yang digunakan. Konversi tersebut dilakukan dengan menggunakan WEKA. Di mana baik ketika digunakan data numerik maupun nominal dapat menghasilkan nilai RAE dan RMSE. Namun, dapat dilihat bahwa hasil evaluasi error dengan RAE dan RMSE menghasilkan nilai error yang berbeda dari data yang sama. Dapat dilihat dari tabel di atas bahwa dataset dengan tipe nominal menghasilkan nilai yang lebih rendah disbanding dataset dengan tipe numerik.

TABEL XXII
PARAMETER YANG MENGHASILKAN ESTIMASI TERBAIK (DATASET CHINA)

No	**Hasil Seleksi Fitur**	**Kode parameter**	**Keterangan**
1	8.899	ID	Kode Proyek Perangkat Lunak
2	8.857	N_effort	Jumlah cost per setiap unit proyek (module)
3	8.2	AFP	Jumlah Baris Koding Hasil AFP(Adjusted Function Points)
4	7.756	Added	Unit Tambahan pada Proyek Software
5	7.542	NPDR_AFP	Jumlah Baris Koding Hasil NPDR_AFP(SLOC generated Normalized productivity delivery rate - Adjusted Function Points)
6	7.532	NPDR_UFP	Jumlah Baris Koding Hasil NPDU_UFP (Normalized productivity delivery rate - UnAdjusted Function Points
7	7.498	PDR_AFP	Jumlah Baris Koding Hasil PDR_AFP(Adjusted Function Points)
8	7.469	PDR_UFP	Jumlah Baris Koding Hasil NPDR_UFP(UnAdjusted Function Points)
9	7.29	Input	Masukan Software
10	6.804	Output	Hasil Software
11	6.16	Enquiry	Data Requirement (spesifikasi Kebutuhuan fungsional dan non fungsional)
12	6.01	File	Jumlah file Coding Proyek
13	4.609	Changed	Unit Perubahan pada Proyek Software
14	4.275	Duration	Durasi Pengerjaan Proyek
15	3.368	Interface	Komponen antar muka Software
16	1.198	Resource	Sumber daya pada sistem contoh: memori dll.
17	1.193	Deleted	Unit yang di buang pada Proyek Software
18	0	Dev.Type	Tipe device yang digunakan (Hardware)

Dari tabel 22 di atas dapat dilihat bahwa estimasi terbaik atau nilai RAE dan RMSE terkecil dihasilkan oleh hasil seleksi fitur menggunakan *information gain* pada dataset china. Pada tabel tersebut terdapat keterangan setiap atribut atau parameter proyek datasetnya.

V. KESIMPULAN

Seleksi Fitur berhasil menurunkan nilai error estimasi (yang diwakilkan oleh nilai RAE dan RMSE). Artinya bahwa semakin rendah nilai error (RAE dan RMSE) maka semakin akurat nilai estimasi yang dihasilkan. Estimasi semakin baik

setelah di lakukan seleksi fitur baik menggunakan *information gain* maupun *mutual information*. Dari nilai error yang dihasilkan maka dapat disimpulkan bahwa dataset yang dihasilkan seleksi fitur dengan metode *information gain* lebih baik disbanding *mutual information* namun, perbedaan keduanya tidak terlalu signifikan. Dari hasil nilai error maka dataset terbaik dalam melakukan estimasi proyek perangkat lunak adalah dataset China. Artinya pemilihan parameter proyek dalam dataset china sangat cocok di hitung untuk melakukan estimasi (*Software Cost Estimation*).

Parameter yang terbaik untuk melakukan *Software Cost Estimation* adalah parameter pada dataset china yang mana rangking tertinggi secara mayoritas diwakilkan oleh perhitungan SLOC (*Source Line Of Code*) atau jumlah baris koding. Jumlah parameter atau jumlah *instance* yang digunakan dalam melakukan *software cost estimation* atau dalam mengestimasi biaya software tidak mempengaruhi hasil estimasi.

Penelitian ini dapat dikembangkan untuk di estimasi menggunakan teknik pembelajaran mesin lainny selain K-*Nearest neighbor*. Dalam melakukan estimasi *software* dapat digunakan model yang sudah dibuat dalam penelitian ini, yakni menggunakan contoh-contoh parameter yang telah diseleksi fitur dan terbukti menghasilkan nilai yang baik.

DAFTAR PUSTAKA

[1] Sharma, M. and Fotedar, N., 2014. Software Effort Estimation with Data Mining Techniques-A Review. *International journal of engineering sciences and research technology*, *3*(3).

[2] Boehm, B.W., 1988. Understanding and controlling software costs. *Journal of Parametrics*, *8*(1), pp.32-68.

[3] Rasywir, E. and Purwarianti, A., 2016. Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin. *Jurnal Cybermatika*, *3*(2).

[4] Adhitya, E.K., Wahono, R.S. and Subagyo, H., 2015. Komparasi Metode Machine Learning dan Metode Non Machine Learning untuk Estimasi Usaha Perangkat Lunak. *Journal of Software Engineering*, *1*(2), pp.109-113.

[5] Yang, Y. and Liu, X., 1999, August. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42-49). ACM.

[6] Yang, Y. and Pedersen, J.O., 1997, July. A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412-420).

[7] Albrecht, A.J. and Gaffney Jr, J.E., 1993, October. Software function, source lines of code and envelopment effort prediction: a software science validation. In *Software engineering metrics I* (pp. 137-154). McGraw-Hill, Inc..

[8] Ratnasari, A., Ardiani, F. and Nurvita, A., 2013. Penentuan Jarak Terpendek dan Jarak Terpendek Alternatif Menggunakan Algoritma Dijkstra Serta Estimasi Waktu Tempuh. *Semantik 2013*, *3*(1), pp.29-34.

[9] Choy, S.K., Tang, M.L. and Tong, C.S., 2011. Image segmentation using fuzzy region competition and spatial/frequency information. *IEEE Transactions on Image Processing*, *20*(6), pp.1473-1484.

[10] Danger, R., Segura-Bedmar, I., Martínez, P. and Rosso, P., 2010. A comparison of machine learning techniques for detection of drug target articles. *Journal of biomedical informatics*, *43*(6), pp.902-913.

[11] Dennis, A. and Wixom, B.H., 2000. System analysis and design: An applied approach. *New York*.

[12] Gary D. Boetticher, 2001, Using Machine Learning to Predict Project Effort: Empirical Case Studies in Data-Starved Domains.

[13] http://www.philadelphia.edu.jo/it/cs/syllabus/731332.pdf

[14] M. Shepperd, 2010, The NAME Project Non- Algorithmic Methods of Estimating, http://dec.bournemouth.ac.uk/staff/decind22/web/NAME.html.

[15] Riyanarto Sarno, Joko Lianto Buliali & Siti Maimunah, 2002, Pengembangan Metode Analogy Untuk Estimasi Biaya Rancang Bangun Perangkat Lunak: Jurnal Makara, Teknologi, vol. 6, no. 2, Agustus

[16] Shepperd, M. and MacDonell, S., 2012. Evaluating prediction systems in software project estimation. *Information and Software Technology*, *54*(8), pp.820-827.