JICN: Jurnal Intelek dan Cendikiawan Nusantara
https://jicnusantara.com/index.php/jicn
Vol : 1 No: 3, Juni - Juli 2024
E-ISSN : 3046-4560

TEKNIK RESAMPLING UNTUK MENINGKATKAN NILAI AKURASI ALGORITMA RANDOM FOREST PADA DATA PREDIKSI KECACATAN PERANGKAT LUNAK

RESAMPLING TECHNIQUE TO INCREASE THE ACCURACY VALUE OF RANDOM FOREST ALGORITHM ON SOFTWARE DEFECT PREDICTION DATA

Sindrawati[1], Dodi Syaripudin[2], Abu Walad[3]
[1,2,3]Fakultas Kesehatan dan Teknik, Universitas Bandung, Bandung, Indonesia
Email: sindrawati@Bandunguniversity.ac.id[1], dodisyaripudin@bandunguniversity.ac.id[2], abuwalad@bandunguniversity.ac.id[3]

Article Info
Article history :
Received : 14-07-2024
Revised : 16-07-2024
Accepted : 20-07-2024
Published : 25-07-2024

Abstract

Software is the main media needed to do their daily activities. Any Defect in the Software often become problematic and need a series of test to reduce it, However, the process itself needs large cost. Therefore, to minimize the tests expenses, it requires data mining research to help the software defect prediction performance. This research uses 12 data sets from NASA Repository which was classified using the Random Forest model by using the Filterisation Technique Resample. The result from this research is an accuracy value and AUC curve AUC whereas the highest accuracy is 99.06% from the PC 2 dataset and concluded that if the Random Forest algorithm use the resample technique then it is the appropriate method to classified Software dataset defect prediction taken from NASA Repository.

Keywords: Software Defect Prediction, Resample, Random Forest, ROC, AUC

Abstrak

Software merupakan perantara terpenting yang dibutuhkan setiap orang untuk menunjang aktivitasnya sehari-hari. Kesalahan perangkat lunak sering kali terjadi dan menjadi kendala, sehingga perlu dilakukan pengujian perangkat lunak untuk mengurangi tingkat kesalahannya, namun proses pengujian kesalahan perangkat lunak memerlukan biaya yang tidak sedikit, untuk meminimalkan biaya pengujian, dilakukan penelitian Data mining diperlukan. untuk memprediksi kesalahan perangkat lunak. Penelitian ini menggunakan 12 kumpulan data arsip NASA yang diklasifikasikan dengan model yang digunakan dalam penelitian ini yaitu model hutan acak yang menggunakan metode resampling filter. Hasil yang diperoleh pada penelitian ini berupa nilai akurasi dan kurva AUC, nilai akurasi tertinggi terdapat pada dataset komputer 2 dengan nilai akurasi sebesar 99,06%, sehingga dapat dikatakan algoritma Random Tree Forest menggunakan teknik Resampling adalah metode yang tepat ketika digunakan untuk mengklasifikasikan kumpulan data perangkat lunak prediksi kegagalan. Arsip NASA.

Kata Kunci: Software Defect Prediction, Resample, Random Forest, ROC, AUC

PENDAHULUAN

Dalam era digital ini, penggunaan teknologi dan algoritma kecerdasan buatan telah menjadi pilihan yang efektif untuk menganalisis data (Raharja et al., 2024). Penggunaan teknologi informasi memungkinkan kita untuk memperoleh dan memproses informasi dengan lebih efisien,

meningkatkan komunikasi dan engoptimalkan pengambilan keputusan(Ramalinda & Raharja, 2024).

Perangkat Lunak (Software) merupakan sekumpulan program dari komputer yang bisa membantu meringankan dari pekerjaan serta dapat disesuaikan dengan yang dibutuhkan oleh user(Raharja et al., 2024), selain itu dapat bertanggung jawab dalam semua sistem yang sedang berjalan pada komputer (Rismayadi et al., 2024). Perangkat Lunak juga dapat dikatakan sebagai sebuah program serta berbagai informasi yang dapat diproses oleh komputer suatu program yang dirancang oleh pengembang untuk bisa diartikan oleh komputer (Ningsih, 2019). Menurut bahasa software adalah kumpulan dari data yang tersimpan dalam komputer serta dapat dilakukan pengaturan oleh komputer, maka dari itu dengan adanya software maka suatu komputer dapat menjalankan perintah dengan baik (Hidayat, Made, & Iswari, 2018).

Sebuah Software dengan bug didalamnya dapat dikatakan ada cacat tersembunyi atau program tersebut tidak berjalan sebagaimana fungsinya (Muchsam et al., 2023). Pengembang perangkat lunak mengembangkan, membuat, dan menggunakan perangkat lunak untuk bisnis dan lembaga pemerintah. Tujuan dari rekayasa perangkat lunak adalah untuk memastikan bahwa sistem yang dikelola oleh perusahaan atau lembaga pemerintah bekerja secara efisien. Meningkatkan kualitas pengujian perangkat lunak memerlukan alat yang membantu memprediksi kesalahan perangkat lunak, namun mengembangkan dan memelihara perangkat lunak yang baik membutuhkan biaya yang mahal(Tiur et al., 2024)

Software Defect Prediction atau memprediksi dari kecacatan sebuah perangkat lunak dapat dimanfaatkan untuk memberikan identifikasi dari suatu modul yang rentan terhadap kecacatan perangkat lunak dan membantu dalam prediksi kesalahan pada perangkat lunak tersebut(Sutisna et al., 2024). Dalam pengecekan kecacatan pada suatu software perlu adanya melakukan proses klasifikasi dengan menggunakan metode klasifikasi yang baik, akan tetapi pada study kasus yang diambil data perlu dilakukan filterisasi agar nilai akurasi dari metode klasifikasi bisa lebih meningkat. Fokus penelitian saat ini adalah: 1. Memprediksi jumlah kecacatan software pada sistem, 2. Menemukan hubungan antara kecacatan ini, dan 3. Mengklasifikasikan kecacatan software. Membangun sistem untuk memprediksi cacat software yang baik dapat mengurangi biaya pengembangan software. (Arar & Ayan, 2015).

Model prediktor kegagalan software telah menjadi subjek penelitian selama lebih dari dua dekade. Sebelum modul software diuji lebih lanjut, model algoritma digunakan untuk memprediksi kesalahan., dari prediksi kecacatan suatu perangkat lunak bisa didapat informasi perangkat lunak mana yang memerlukan penanganan lebih intens. Bagian yang bertugas untuk pengecekan software dapat berdasarkan hasil pengujian model algoritma, mengatur waktu dan biaya agar lebih cepat dan murah. Prediksi cacat membutuhkan lebih banyak fokus daripada prediksi tidak cacat. (Wattiheluw, Rochimah, & Fatichah, 2019).

Data metrik perangkat lunak sangat populer dalam pengembangan model prediksi cacat software. Dataset NASA berasal dari dua sumber, yaitu repository NASA MDP (Metrics Data Program) dan repository Predictor Models in Software Engineering (PROMISE) (Raharja, 2024).

Dataset NASA mudah diperoleh dan kinerja metode yang digunakan dapat dibandingkan dengan dataset lain.

Bedasarkan penelitian beberapa algoritma yang sering digunakan untuk pengklasifikasian seperti algoritma C4.5, algoritma Decision Tree, algoritma Linear Regression, algoritma Logistic Regression (LR), algoritma Naïve Bayes (NB), algoritma Neural Network (NN), algoritma Random Forest (RF) dan algoritma Support Vector Machine (SVM) menjadi fokus topik penelitian (Bowes et al., 2011). Random forest adalah algoritma klasifikasi yang menggunakan ensemble learning. Random forest didasarkan pada sebuah ide untuk membentuk suatu kumpulan dari decision tree dengan variansi yang dapat diatur (Breiman, 2001). Untuk meningkatkan kinerja, gabungan adalah metode divide and conquer. Prinsip utama metode kelompok adalah bahwa kelompok "pelajar yang lemah" dapat digabungkan untuk membentuk sebuah "pelajar yang kuat". Random Forest memiliki waktu kerja yang cukup cepat dan mampu menangani data yang tidak seimbang dan tidak lengkap. Kelemahan dari regresi adalah mereka tidak dapat memprediksi nilai di luar jangkauan mereka pada data pelatihan, dan mereka mungkin melakukan overfitting pada data yang cukup noise.(Imaduddin, 2014). Random forest memberikan nilai rata-rata untuk menemukan titik balance pada data-data tersebut. Random Forest tahan terhadap gangguan yang terdapat dalam data. (Imaduddin, 2014).

Random Forest Merupakan Algoritma training dengan menggunakan penggabungan bootstrap (bagging). Satu set data latihan dikumpulkan dan dimasukkan ke dalam suatu pohon untuk memulai proses latihan. Setiap kali sebuah node dipecah, pemilihan atribut dilakukan secara acak. Baging melakukan pemilihan sample berulang dengan penggantian. Jumlah data latih untuk setiap pohon akan sama. Nilai treshold suatu node dapat dihitung dengan menggunakan indeks Gini. (Han & Kamber, 2006). Berdasarkan kinerja matrik, *Random Forest* adalah classifier terbaik di antara pengklasifikasian lainnya (Fiandra et al., 2017).

Software matrix yang digunakan oleh NASA , untuk memprediksi cacat perangkat lunak, secara umum yang menggunakan dataset NASA yang mengalami *imbalanced class* (Khoshgoftaar et al., 2014). Ada dua pendekatan umum yang dipakai untuk menangani permasalahan dataset tidak seimbangan, yaitu pendekatan level data dan pendekatan level algoritma (Zhang et al., 2008). Menangani dataset berdimensi tinggi selain seleksi fitur (*feature selection*), dapat juga menggunakan teknik resampling. Peningkatkan data di dalam kelas minoritas dapat meningkatkan kemampuan algoritma *machine learning* menjadi lebih baik, karena bisa mengenali sampel kelas minoritas dari sampel mayoritas (Thanathamathee & Lursinsap, 2013).

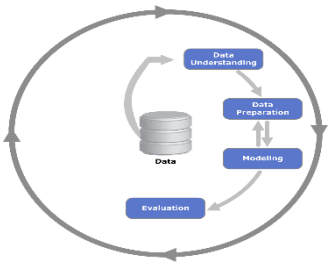
Teknik Resample Ini adalah cara paling populer untuk mengatasi masalah ini. Ada tiga cara utama untuk mengatasi masalah underclassification: yaitu dengan menggunakan sampel besar untuk kelas minoritas, sampel kecil untuk kelas mayoritas, atau metode hybrid menurut keduanya. Modelnya sendiri merupakan cara untuk mengubah sebaran kelas minoritas saat melatih data untuk algoritma pembelajaran mesin agar kelas tersebut tidak menjadi kelas minoritas. Metode pemodelan telah terbukti dapat menyelesaikan permasalahan klasifikasi yang tidak pasti (Jayadi et al., 2024). Random sampling (ROS) adalah cara paling sederhana untuk menangani kelas minoritas dengan merangkum kelas-kelas selama proses pengambilan sampel. Cara pengambilan sampel dengan

metode ROS adalah dengan mengembalikan kelas terbaik dengan fungsi Random Class Matching (Ganganwar, 2012). Karena metode ROS ini mengembalikan kelas yang baik di kelas minoritas, hal ini menimbulkan risiko redundansi. Random subsampling mirip dengan metode random super sampling dalam menghitung selisih antara kelas mayoritas dan kelas minoritas. Kemudian ulangi perbedaan antara layer utama dan layer kecil beberapa kali. Selama operasi, sebagian besar kelas dihapus menurut subkelasnya (Rahayu et al., 2024)

Pada penelitian ini yang akan di lakukan adalah penerapan *resample* untuk penyelesaian ketidakseimbangan kelas pada *Random Forest* untuk prediksi cacat perangkat lunak, sehingga dapat menghasilkan kinerja yang baik terhadap data set yang seimbang. Melihat dari permasalahan tersebut maka dalam penelitian ini akan membahas tentang "Teknik Resampling Untuk Meningkatkan Nilai Akurasi Algoritma *Random Forest* Pada Data *Software Defect Prediction*".

METODE PENELITIAN

Pada tahapan ini, menjelaskan metodologi penelitian secara keseluruhan dengan menggunakan metode CRISP-DM. Berikut merupakan tahapan yang dilalukan pada penelitian ini seperti yang terlihat pada gambar 3.1.



Gambar 3.1 Metode CRISP-DM

Instrument Penelitian

Pengumpulan data dalam penelitian tindakan kelas ini dengan menggunakan instrumen utama dan instrumen penunjang. Instrumen utama adalah peneliti bahwa peneliti adalah orang yang paling mengetahui seluruh data dan cara menyikapinya.(Lestari , 2019).

Pada penelitian ini digunakan beberapa instrumen penelitian, antara lain sebagai berikut :

1. Perangkat Lunak (Software) Perangkat lunak yang akan digunakan dalam penelitian ini adalah software weka 3.8.
2. Dataset Penelitian ini menggunakan dataset dari nasa repository sebanyak 12 dataset tahun 2018 dengan 12 data set.

Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data public yang di ambil dari dataset nasa repository.

Metode Analisa Data

Pengolahan Data Awal

Pada tahap ini merupakan tahap untuk memastikan dataset NASA repository yang dipilih telah layak untuk dilakukan proses pengolahan.

Data Understanding

Pada Tahapan ini menerangkan tentang data yang di ambil yaitu data dari nasa repository sebanyak 12 Data set yang berisi tentang data kecacatan perangkat lunak.

Data Preparation

Pada tahapan ini menerangkan tentang data set yang di lakukan proses filtering dengan menggunakan Random Over-Sampling (ROS) dan Random Under-Sampling (RUS) serta Teknik Sampling Minority Over-Sampling (SMOTE) adalah teknik resampling yang menggunakan data latih untuk memperbaiki kecondongan distribusi kelas.

Modeling

Pada tahapan ini menerangkan tentang model yang di gunakan dalam penelitian ini yaitu dilakukan beberapa kali pengujian menggunakan beberapa model yang terdapat pada tools weka dan di dapatkan 1 model yaitu model random forest karena mendapatkan nilai akurasi yang tinggi pada 12 data set yang diujikan .

Evaluation

Pada tahapan ini setelah model di terapkan dilakukan evaluasi pada ke 12 data set sehingga terdapat 1 data set yang nilai akurasi dan AUC nya lebih tinggi dari pada dataset lainnya .

HASIL DAN PEMBAHASAN

Dataset

Dalam penelitian ini, kami mengumpulkan data sekunder dari Metrics Data Program (MDP) Badan Penerbangan dan Antariksa Nasional (NASA) dan perangkat lunak Matrix, kumpulan data yang digunakan oleh peneliti yang melakukan penelitian teknologi komputer [3]. Pusat Data MDP NASA didedikasikan untuk membahas studi kesalahan komputer dan perangkat lunak. Data NASA dapat ditemukan di arsip PROMISE dan MDP. Data NASA dari MDP digunakan oleh sebagian besar peneliti. karena Martin Shepperd [14] memperbaikinya dengan menghapus data yang hilang atau tidak dilaporkan. Kumpulan data penyimpanan MDP NASA ditunjukkan pada Tabel 4.1

.

Tabel 4.1 Dataset NASA Repository

	Dataset	Keterangan
1	CM1 (Attribute : 38 Modul : 327 Donor: Tim Menzies (tim@barmag.net) Date: December 2, 2004 Sources: *Creators:* NASA, then the NASA Metrics Data Program, http://mdp.ivv.nasa.gov.)	adalah instrumen pesawat ruang angkasa *nasa* (pengumpulan dan pemrosesan data) yang ditulis dalam "C". Di berbagai waktu, peneliti telah menegosiasikan akses ke kode sumber CM. Oleh karena itu, agak lebih banyak dipelajari bahwa beberapa set data *nasa* lainnya. Misalnya, model *UML* untuk CM1 direkayasa balik dari artefak yang disediakan untuk *West Virginia University*.
2	JM1 (Attribute: 22 Modul: 7782) dan MW1 (Attribute: 38 Modul: 253) (Donor: Tim Menzies (tim@barmag.net) Date: December 2, 2004 Sources: *Creators:* NASA, then the NASA Metrics Data Program, http://mdp.ivv.nasa.gov.)	adalah set data *promise* yang dibuat tersedia untuk umum untuk mendorong model rekayasa perangkat lunak yang dapat diulang, diverifikasi, dapat disangkal, dan / atau dapat diprediksi.
3	PC1(Attribute: 38) PC2(Attribute: 37) PC3(Attribute: 38) PC4(Attribute: 38) PC5(Attribute: 39) MC1 (Attribute: 39) dan MC2(Attribute: 39) Donor:Tim Menzies (tim@barmag.net) Date: December 2, 2004 Sources: *Creators:* NASA, then the NASA Metrics Data Program. http://mdp.ivv.nasa.gov	adalah set data yang menerangkan tentang kecacatan *metrik nasa* .Data dari perangkat lunak penerbangan untuk satelit yang mengorbit bumi. Data berasal dari *McCabe* dan *Halstead* fitur ekstraktor kode sumber .Fitur-fitur ini didefinisikan pada tahun 70-an dalam upaya untuk menandai fitur kode secara objektif yang terkait dengan kualitas perangkat lunak.

4	KC2 (Attribute: 22 Modul: 522) dan KC3 (Attribue: 40 Modul: 194) Donor: (Tim Menzies (tim@barmag.net) Date: December 2, 2004 Sources: *Creators:* NASA, then the NASA Metrics Data Program, http://mdp.ivv.nasa.gov.)	adalah set data yang menerangkan tentang kecacaan *metrik nasa*. Data dari perangkat lunak penerbangan untuk satelit yang mengorbit bumi. Data berasal dari *McCabe* dan *Halstead* fitur ekstraktor kode sumber .Fitur-fitur ini didefinisikan pada tahun 70-an dalam upaya untuk menandai fitur kode secara objektif yang terkait dengan kualitas perangkat lunak. Pengertian KC2,KC3 dan PC1,PC2,PC3,PC4,PC5,MC1,MC2 terdapat persamaan akan tetapi ada perbedaan di jumlah atribut. Atribut pada PC1,PC2,PC3,PC4,PC5,MC1,MC2 23 atribut akan tetapi di KC2,KC3 terdapat 22 atribut.

Sumber: (https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/, 2004)

1) Activity Diagram

Hitungan

Pengujian metode dimulai dengan membagi dataset menjadi dua bagian; bagian pertama digunakan sebagai instruksi data dan bagian kedua digunakan sebagai pengujian data untuk mevalidasi model. (Bowes et al., 2011).

Uji akurasi dengan cross validation merupakan metode yang digunakan untuk menguji tingkat akurasi pada pohon keputusan apakah sudah akurat atau belum.

Perhitungan:

1. Data Training : 327 Dataset CM1.
2. Nilai alpha : 60% berarti 196 data training dan 40% atau 131 data testing.

Langkah selanjutnya pengujian sebanyak ‘K’ kali. Nilai K di pilih secara random, K= 10. Kemudian pengujian sebanyak 10x dengan data training CM1 di data tersebut, Hasil perhitungan dapat dilihat pada Tabel 4.2 :

Tabel 4.2 Ilustrasi 10-Fold Cross Validation dan Nilai Akurasi Algoritma Random Forest Pada Dataset CM1

Pengujian data (iterasi)	Banyak data	Pengujian Dataset CM1 (Testing - Training)										Akurasi
1	196-131	Test	train	train	train	train	train	train	train	train	train	88%
2	196-131	train	Test	train	train	train	train	train	train	train	train	86%
3	196-131	train	train	Test	train	train	train	train	train	train	train	86%
4	196-131	train	train	train	Test	train	train	train	train	train	train	86%
5	196-131	train	train	train	train	Test	train	train	train	train	train	86%
6	196-131	train	train	train	train	train	Test	train	train	train	train	88%
7	196-131	train	train	train	train	train	train	Test	train	train	train	87%
8	196-131	train	train	train	train	train	train	train	Test	train	train	84%
9	196-131	train	train	train	train	train	train	train	train	Test	train	87%
10	196-131	train	train	train	train	train	train	train	train	train	Test	85%
		Rata-Rata										86%

Berdasarkan Tabel 4.2 ditunjukkan bahwa nilai fold yang di gunakan adalah 10-fold cross validation. Berikut langkah-langkah pengujian data dengan 10-fold cross validation.

Maka nilai x validation untuk akurasi yaitu x = 86%

Percobaan data baru di setiap pengujian dari 1 hingga 10.
Hasil pengujian :

1.Uji 1> di terima
2.Uji 2> di terima
3.Uji 3> di terima
4..Uji 4> di terima
5. Uji 5> di terima
6. Uji 6> di terima
7. Uji 7> di terima
8. Uji 8> di terima
9. Uji 9> di terima
10. Uji 10> di terima

Dataset yang digunakan dibagi menjadi 10 bagian, yaitu D1, D2, D3, D4, D5, D6, D7, D8, D9 dan D10. Dt, t = (1,2,3,4,5,6,7,8,9,10) digunakan sebagai data testing dan dataset yang lainnya sebagai data training.

Tingkat akurasi di hitung pada setiap iterasi (iterasi-1, iterasi-2, iterasi-3, iterasi-4, iterasi-5, iterasi-6, iterasi-7, iterasi-8, iterasi-9, dan iterasi-10), kemudian dihitung rata-rata tingkat akurasi dari seluruh iterasi untuk mendapatkan tingkat akurasi data keseluruhan dapat dilihat pada Tabel 4.2 akurasi. Selanjutnya dilakukan langkah evaluasi menggunakan area under curve (AUC) untuk mengukur hasil akurasi berdasarkan performa model prediksi. Hasil akurasi dapat diperiksa dengan membandingkan klasifikasi menggunakan kurva Receiver Operating Characteristic (ROC) dari hasil matriks konfusi. ROC menghasilkan dua garis berupa true positif pada garis vertikal dan false positif pada garis horizontal (Defiyanti, 2013).

Kurva ROC merupakan plot antara sensitivitas (tingkat positif sebenarnya) pada sumbu Y dan spesifisitas 1 (tingkat positif palsu) pada sumbu X. Kurva ROC ini menunjukkan adanya antarmuka antara sumbu Y dan sumbu Y. sumbu. Pengukuran akurasi melalui matriks kebingungan. Ada beberapa metode yang digunakan untuk menentukan presisi, sensitivitas (recall), spesifisitas, nilai prediksi positif (PPV) atau presisi, nilai prediksi negatif (NPV) atau FPrate, F-measure atau G-means, APER, dan I-APER Perhitungannya adalah sebagai berikut (Gorunescu, 2011):
Rumus Confusion Matrix :

1.Accuracy =(TP+TN)/(TP+TN+FP+FN)
2.Sensitivity(Recall) = TPrate=TP/(TP+FN)
3.Spesificity = TNrate=TN/(TN+FP)
4.FPrate=FP/(FP+TN)
5.Precision=TP/(TP+FP)
6.F-Measure=(2*P*R)/((P+R))
7.G-mean=√sensitivity*specificity
8.Apparent Error Rate(APER)=(FP+FN)/N
9.Total Accuracy Rate (I-APER)=(TP+TN)/N

JICN: Jurnal Intelek dan Cendikiawan Nusantara
https://jicnusantara.com/index.php/jicn
Vol : 1 No: 3, Juni - Juli 2024
E-ISSN : 3046-4560

Tabel 4.3 Hasil Experimen Random Forest

Dataset	TP	TN	FP	FN	Akurasi	Recall	Specificity	Precision	Fprate	F-Measure	G-mean	AUC
CM1	2	278	7	40	85,63%	0,048	0,975	0,222	0,025	0,078	0,2163	0,77
JM1	342	5833	277	1338	79,35%	0,205	0,955	0,553	0,045	0,299	0,4424	0,701
KC2	384	51	56	31	83,33%	0,925	0,477	0,873	0,523	0,898	0,6642	0,825
KC3	5	153	5	31	81,44%	0,139	0,968	0,5	0,032	0,217	0,3668	0,736
MC1	9	1932	10	37	97,64%	0,196	0,995	0,474	0,005	0,277	0,4416	0,883
MC2	17	71	10	27	70,40%	0,386	0,877	0,63	0,123	0,478	0,5818	0,717
MW1	5	217	9	22	87,75%	0,185	0,96	0,357	0,04	0,243	0,4214	0,716
PC1	15	637	7	46	92,48%	0,246	0,989	0,682	0,011	0,361	0,4932	0,878
PC2	0	729	0	16	97,85%	0	1	0	0	0,00	0	0,773
PC3	15	920	23	119	86,82%	0,112	0,976	0,395	0,024	0,174	0,3306	0,831
PC4	1256	67	111	24	90,74%	0,981	0,376	0,919	0,624	0,949	0,6073	0,945
PC5	16561	224	292	109	97,67%	0,993	0,434	0,983	0,566	0,988	0,6564	0,977

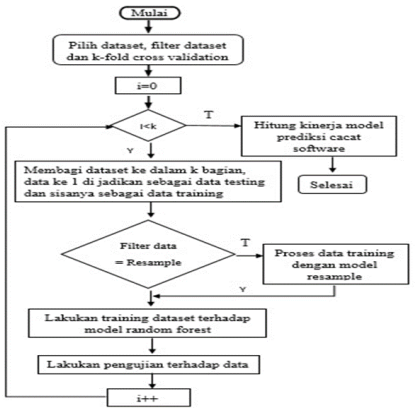
Eksperimen ini menggunakan dua belas dataset dari Perpustakaan NASA (CM1, JM1, KC2, KC3, MC1, MC2, MW1, PC1, PC2, PC3, PC4, dan PC5). Metode pengklasifikasi Random Forest digunakan untuk menguji informasi. Informasi yang diuji termasuk akurasi (recall), sensitivitas (recall), spesifisitas, nilai positif prediktor (PPV) atau ketepatan, nilai negatif prediktor (NPV) atau FPrate, F-Measure, G-Mean, dan AUC. Rata-rata akurasi pada dua belas dataset adalah 87,59%, dan AUC rata-rata adalah 0,812.Sedangan pada Tabel 4.6 ditunjukan hasil eksperimen metode Random Forest dengan Resample untuk 12 dataset NASA Repository.

Tabel 4.4 hasil Pengukuran Random Forest dan Resample

Dataset	TP	TN	FP	FN	Akurasi	Recall	Specificity	Precision	Fprate	F-Measure	G-mean	AUC
CM1	28	284	1	14	95%	0,667	0,996	0,966	0,004	0,789	0,815	0,95
JM1	1124	5968	142	548	91%	0,672	0,977	0,888	0,023	0,795	0,8102	0,926
KC2	407	79	28	8	93%	0,981	0,738	0,936	0,262	0,958	0,8508	0,938
KC3	22	153	5	14	90%	0,611	0,948	0,815	0,052	0,698	0,749	0,895
MC1	33	1957	5	13	99%	0,717	0,997	0,868	0,003	0,786	0,8454	0,996
MC2	20	74	7	16	81,60%	0,636	0,914	0,8	0,086	0,709	0,7624	0,913
MW1	16	224	2	11	95%	0,593	0,991	0,889	0,009	0,711	0,7665	0,917
PC1	46	640	4	15	97,31%	0,754	0,994	0,92	0,006	0,829	0,8657	0,971
PC2	11	727	2	5	99%	0,688	0,997	0,846	0,003	0,759	0,8281	0,987
PC3	87	930	13	47	94%	0,649	0,986	0,87	0,014	0,744	0,7999	0,946
PC4	1172	116	62	8	95%	0,994	0,652	0,954	0,348	0,973	0,805	0,982
PC5	16628	391	125	41	99%	0,998	0,758	0,993	0,242	0,995	0,8697	0,982

Hasil eksperimen pada Tabel 4.4 menunjukan rata-rata akurasi pada 10 dataset adalah 94,20% dan rata-rata AUC sebesar 0,9445.

Flowchart



JICN: Jurnal Intelek dan Cendikiawan Nusantara
https://jicnusantara.com/index.php/jicn
Vol : 1 No: 3, Juni - Juli 2024
E-ISSN : 3046-4560

Weka

1. Pengukuran penelitian

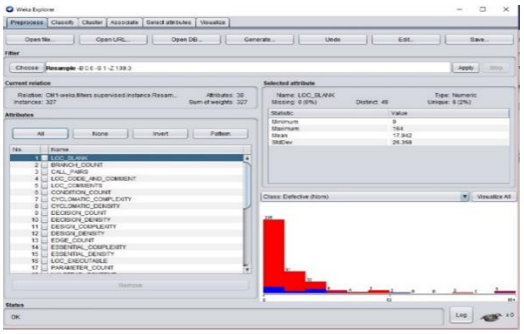
Pengukuran penelitian ini membahas tentang hasil penelitian dan pengujian model yang dipakai dalam penelitian sehingga hasilnya sesuai penelitian yang dilakukan.

Hasil dari penelitian ini menguji keakuratan prediksi kecacatan dari perangkat lunak dengan menggunakan algoritma Random Forest beserta filterisasi Resample, dengan tujuan untuk mengetahui dan menguji tingkat ke akuratan dari algoritma yang di gunakan sehingga kecacatan suatu software bisa di prediksi dengan baik. Data yang dianalisa adalah data set dari nasa repository tentang kecacatan perangkat lunak.

2. Pengujian Model

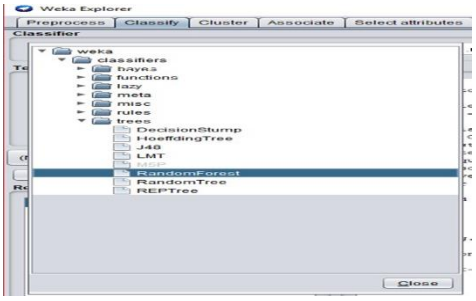
Model yang dibuat dari data uji diuji menggunakan metode 10 cross validation dimana data dibagi secara acak menjadi 10 bagian. Cara paling akurat untuk mendapatkan hasil adalah dengan menentukan kesalahannya. Setelah menghitung tingkat kesalahan setiap kategori, maka dihitung rata-rata keseluruhannya (Fauzi dan Tukiyat, 2019)

Setelah dilakukan klasifikasi model data, maka tahap selanjutnya melakukan pengujian data untuk memprediksi akurasi data uji.



Gambar 4.2 Input Dataset dan Resample

Pada gambar 4.2 merupakan proses dimana dataset di input dan kemudian dilakukan filterisasi menggunakan metode resample dengan tujuan metode tersebut dapat menghilangkat data yang di anggap noise.



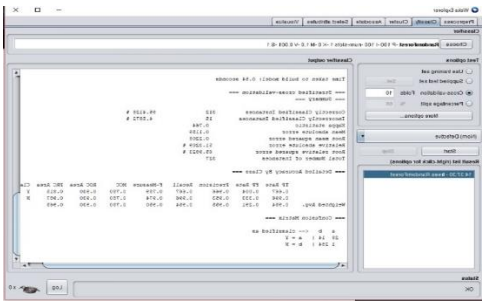
Gambar 4.3 Input Algoritma

JICN: Jurnal Intelek dan Cendikiawan Nusantara
https://jicnusantara.com/index.php/jicn
Vol : 1 No: 3, Juni - Juli 2024
E-ISSN : 3046-4560

Pada gambar 4.3 merupakan proses dimana data yang telah di lakukan filterisasi mengggunakan metode resample dilakukan klasifikasi menggunakan model random forest. Hal tersebut dilakukan karena model random forest merupakan model yang memiliki tingkat akurasi paling tinggi jika di banding kan dengan menggunkan metode lainnya.

Gambar 4.4 Menentukan Cross-validation

Pada Gambar 4.4 merupakan proses dimana dilakukan testing dari algoritma random forest dengan menggunakan cross validation sebanyak 10 kali. Hal tersebut di lakukan sesuai dengan penelitian penelitian yang sebelumnya telah dilakukan dan didapatkan bahwa nilai cross validation 10 itu merupakan jumlah testing yang paling baik.

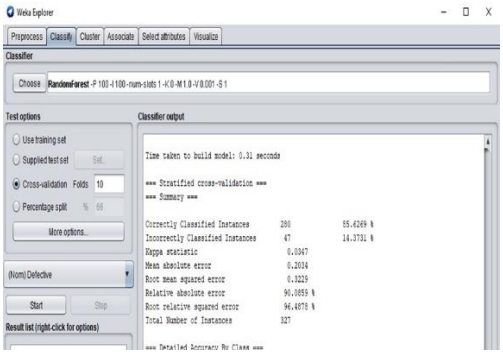


Gambar 4.5 Pengujian Algoritma

Pada gambar 4.5 merupakan hasil dari pengujian yang telah dilakukan sehingga dapat dikatakan bahwa dengan menggunakan teknik resample dan algoritma random forest memiliki nilai akurasi sebesar 95,4128% dengan nilai ROC sebesar 0,930. Hal tersebut menunjukan bahwa performa dari algoritma random forest dengan teknik filterisasi resample jika di terapkan pada data set software defect prediction yang diambil dari data nasa repository merupakan algoritma yang tepat.

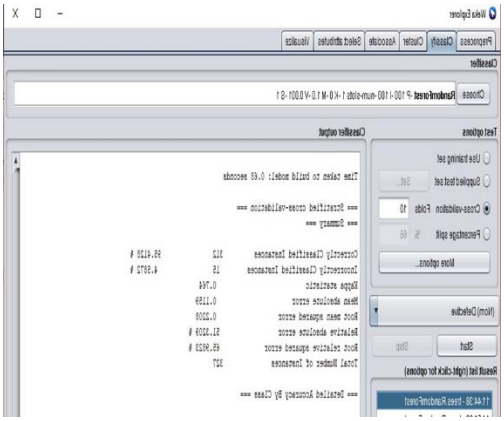
Kemudian masukkan nilai yang ada di dalam confusion matrix ke dalam persamaan di atas ke dalam algoritma Random Forest, sehingga akan menghasilkan nilai seperti di bawah ini:

JICN: Jurnal Intelek dan Cendikiawan Nusantara
https://jicnusantara.com/index.php/jicn
Vol : 1 No: 3, Juni - Juli 2024
E-ISSN : 3046-4560



Gambar 4.8. Confusion Matrix menentukan accuracy dengan random forest

Pada Gambar 4.8. untuk algoritma Random Forest menghasilkan nilai accuracy 85,62 %.



Gambar 4.9. Confusion Matrix menentukan accuracy dengan random forest filterisasi Resample.

Pada Gambar 4.9. Hasil pengujian confusion matrix di atas diketahui menggunakan model algoritma random forest filterisasi resample memiliki akurasi 95,41% dibandingkan model algoritma Random Forest tanpa resample hanya 85,42%. tingkat akurasi yang lebih tinggi dibandingkan algoritma random forest sebesar 9,78%.

KESIMPULAN

Dari penelitian yang telah dilakukan menunjukan rata-rata akurasi pada 12 dataset dengan menggabungan metode filterisasi *Resample* dengan model Random Forest adalah 94,20% dan rata-rata *AUC* sebesar 0,9445. Hasil eksperimen pada penelitian ini mendapatkan nilai akurasi sebesar 99,06% pada data set PC2 dengan model *RF+Resample*, Mengalami peningkatan sebesar 1,21% dari *RF* tanpa *Resample*. Dan Hasil *AUC* sebesar 0.992 pada data set PC5 untuk model *RF+Resample*. Metode *Random Forest* filterisasi *Resample* merupakan metode yang cukup baik dalam pengklasifikasian data *mining*. Hal ini dikarenakan algoritma tersebut dapat menghasilkan nilai akurasi yang cukup tinggi untuk ke 12 dataset sofware defect prediction NASA Repository CM1 95,41%, JM1 91,13%, KC2 93,10%, KC3 90,21%, MC1 99,09%, MC2 81,60%, MW1 94,86%, PC1 97,31%, PC2 99,06%, PC3 94,43%, PC4 95,20%, dan PC5 99,03%.

Saran-saran

1. Pada penelitian berikutnya dapat dilakukan eksperimen menggunakan metode *Random Forest* dengan penerapan algoritma optimasi seleksi fitur yang lain untuk meningkatkan performa terbaik dan mampu meningkatkan nilai akurasi yang tinggi.
2. Penggabungan beberapa algoritma supaya mendapatkan hasil yang cukup tinggi dengan menambahkan seleksi fitur yang mempunyai tingkat akurasi yang tinggi.
3. Dibuatkannya sebuah program atau aplikasi untuk mengimplementasikan metode *Random Forest* tersebut.

DAFTAR PUSTAKA

Bowes, D., Gray, D., Hall, T., Counsell, S., & Beecham, S. (2011). A Systematic Review of Fault Prediction Performance in Software Engineering. *IEEE Transactions on Software Engineering*, *38*(6), 1276–1304.

Breiman, L. (2001). (impo)Random forests(book). *Machine learning*, 5–32. https://doi.org/10.1023/A:1010933404324

Fiandra, Y. A., Defit, S., & Yuhandri, Y. (2017). Penerapan Algoritma C4.5 untuk Klasifikasi Data Rekam Medis berdasarkan International Classification Diseases (ICD-10). *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *1*(2), 82. https://doi.org/10.29207/resti.v1i2.48

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*.

Imaduddin, A. (2014). *Pengenalan Karakter Huruf Hangul Korea*. *1*(1), 755–763.

Jayadi, J., Raharja, A. R., Pramudianto, A., & Muchsam, Y. (2024). *Application of Naïve Bayes Classifier Algorithm for Classification of Scholarship Recipients at SMA PGRI 2 Bandung*. *13*(2), 33–41.

Khoshgoftaar, T. M., Gao, K., Napolitano, A., & Wald, R. (2014). A comparative study of iterative and non-iterative feature selection techniques for software defect prediction. *Information Systems Frontiers*, *16*(5), 801–822. https://doi.org/10.1007/s10796-013-9430-0

Muchsam, Y., Sucipto, B., Rismawati, R., Rusdianti, I. S., & Raharja, A. R. (2023). Forming the Character of a Physically Healthy Young Generation Through Military Education. *TGO Journal of Community Development*, *1*(2), 90–95. https://doi.org/10.56070/jcd.2023.015

Raharja, A. R. (2024). *Keamanan Jaringan*. PENERBIT KBM INDONESIA.

Raharja, A. R., Setiyono, R., & Hariyanti, I. (2024). IMPLEMENTASI APLIKASI SURFACE ROUGHNESS TESTER ATAU ALAT UKUR KEKASARAN PERMUKAAN JALAN MENGGUNAKAN C# DAN ARDUINO. *Media Informatika*, *23*(1), 1–9. https://doi.org/10.37595/mediainfo.v23i1.206

Rahayu, T., Yayat, E., & Raharja, A. R. (2024). *Analysis of Storage Spaces to Support the Health Service System at Santosa Hospital Bandung Central in 2021*. 19–26.

Ramalinda, D., & Raharja, A. R. (2024). DECISION SUPPORT SYSTEM FOR SELECTING RECIPIENTS OF HOME RENOVATION ASSISTANCE USING THE TOPSIS METHOD. *International Journal of …*, *42*(1), 17–24. https://jicnusantara.com/index.php/jicn/article/view/535

Rismayadi, A. A., Wiguna, W., Muchsam, Y., Rumaisa, F., Jayadi, Pramudianto, A., & Raharja, A. R. (2024). PEMBELAJARAN C#. In *Mafy Media Literasi*.

Sutisna, T., Raharja, A. R., Hariyadi, E., Hafizh, V., & Putra, C. (2024). Penggunaan Computer Vision untuk Menghitung Jumlah Kendaraan dengan Menggunakan Metode SSD (Single Shoot Detector). *Journal Of Social Science Research Volume*, *4*, 6060–6067. https://doi.org/10.31004/innovative.v4i2.10071

Thanathamathee, P., & Lursinsap, C. (2013). Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. In *Pattern Recognition Letters* (Vol. 34, Nomor 12). Elsevier B.V. https://doi.org/10.1016/j.patrec.2013.04.019

Tiur, M., Setiatin, S., Ramalinda, D., & Raharja, A. R. (2024). Analysis of Quality Dimensions on The Level of Satisfaction of Health Services in The Covid-19 Pandemic Era (at Cikembar Health Center in 2020). *Journal of Student Collaboration Research*, *1*(1), 30–35.

Zhang, Z. Z., Chen, Q., Ke, S. F., Wu, Y. J., Qi, F., & Zhang, Y. P. (2008). Ranking potential customers based on group-ensemble. *International Journal of Data Warehousing and Mining*, *4*(2), 79–89. https://doi.org/10.4018/jdwm.2008040109