Cenderawasih : Journal of Statistics and Data Science
E-ISSN: 3026-0981

Perbandingan Feature Selection antara *Backward Method* dan *Stepwise Method* pada Dataset *Primary Tumor*

FRISMA AYU WIDYARIMBI[1], JEANLI RISKA ARSANI[2], NICEA ROONA PARANOAN[3]

[1,2,3)] Program Studi Statistika Fakultas MIPA Universitas Cenderawasih Jayapura, Indonesia
e-mail: frismaayuw@gmail.com[1], jeanliriskaaa@gmail.com[2]

ABSTRAK

Feature selection dalam data mining bertujuan mengidentifikasi dan mempertahankan fitur-fitur relevan, meningkatkan kinerja model, serta mempermudah interpretasi, yang menghasilkan model yang efisien dan efektif untuk pengambilan keputusan lebih baik. Penelitian ini bertujuan untuk membandingkan kinerja dua *Feature Selection*, yaitu antara *Backward Method* dan *Stepwise Method*, pada dataset *Primary Tumor*. Pada dataset ini memberikan informasi tentang berbagai jenis tumor primer yang didiagnosis pada pasien di berbagai lokasi dalam tubuh. Metode yang digunakan untuk mengukur kinerja dua *feature selection* pada penelitian ini adalah menggunakan nilai *accuracy*. Nilai *accuracy* yang lebih tinggi menunjukkan bahwa model yang dikembangkan dengan fitur-fitur yang dipilih melalui metode *backward* atau metode *stepwise* memiliki kemampuan yang lebih baik dalam memprediksi data uji atau memodelkan hubungan dalam dataset. Pada penelitian ini, didapatkan hasil bahwa *accuracy* dari metode *backward* cenderung lebih tinggi yaitu sebesar 66.774 (0.021) sedangkan *accuracy* dari metode *stepwise* adalah sebesar 65.806 (0.008). Hal ini menunjukkan bahwa dengan menggunakan metode backward cenderung lebih baik dalam memprediksi variabel dependen dibandingkan dengan model yang dibangun menggunakan metode *stepwise* dalam konteks data yang digunakan.

Kata Kunci: Feature Selection, Backward Method, Stepwise Method, Primary Tumor, Accuracy

1. PENDAHULUAN

Di era digital yang berkembang pesat, produksi dan pengumpulan data telah mencapai proporsi yang belum pernah terjadi sebelumnya. Organisasi dan bisnis dihadapkan dengan semakin banyaknya data dari berbagai sumber, termasuk transaksi bisnis, sensor, media sosial, perangkat seluler, dan banyak lagi. Dalam konteks ini, *data mining* sebagai metode analisis berperan penting dalam mengubah data menjadi wawasan yang berharga. Integrasi teknologi Internet of Things (IoT) dalam sistem informasi kesehatan memungkinkan pemantauan kesehatan pasien secara real-time, yang berkontribusi pada pengumpulan data kesehatan yang lebih banyak dan akurat, serta pengambilan keputusan klinis yang lebih baik (Arie, 2023).

Dengan adanya sistem informasi kesehatan yang terintegrasi dan terhubung, data kesehatan dapat diolah dan dianalisis untuk menghasilkan model klasifikasi yang dapat mendukung pengambilan keputusan klinis yang berbasis data, mempercepat diagnosa, dan meningkatkan efektivitas pengobatan. Analisis data kesehatan dalam skala besar (big data) juga memberikan informasi yang lebih akurat tentang pola penyakit, kinerja sistem kesehatan, dan faktor-faktor risiko kesehatan yang dapat membantu penyusunan kebijakan dan pengambilan keputusan klinis yang lebih baik (Arie, 2023).

Namun, salah satu tantangan utama dalam analisis data medis adalah mengelola jumlah fitur yang besar, yang seringkali tidak semuanya relevan untuk tujuan analisis tertentu. Pemilihan fitur adalah teknik yang dapat membantu mengidentifikasi fitur-fitur yang paling penting dalam suatu dataset, yang pada gilirannya dapat menghasilkan model yang lebih efisien, mengurangi *overfitting*, dan meningkatkan interpretasi hasil. Teknik ini sangat penting dalam konteks data medis, di mana variabel yang tidak relevan dapat menyebabkan model yang kompleks dan sulit diinterpretasikan, serta meningkatkan risiko kesalahan dalam pengambilan keputusan klinis (Leto et al., 2023).

Dalam data medis, di mana keakuratan dan keandalan adalah sangat penting, pemilihan fitur yang efektif dapat membantu dalam menghindari kesalahan yang bisa berakibat fatal dan memastikan bahwa keputusan klinis didasarkan pada informasi yang paling relevan dan akurat. Selain itu, pemilihan fitur juga dapat membantu dalam mengatasi tantangan yang berkaitan dengan dimensi data yang tinggi, yang sering ditemui dalam dataset medis. Pemilihan fitur tidak hanya menguntungkan dari segi pembangunan model, tetapi juga dari segi interpretasi hasil. Fitur yang lebih sedikit dan lebih relevan memudahkan praktisi medis untuk memahami bagaimana model membuat prediksi, yang sangat penting dalam konteks medis di mana penjelasan yang jelas dan transparan diperlukan untuk kepercayaan dan adopsi oleh profesional kesehatan. Oleh karena itu, pemilihan fitur menjadi langkah krusial dalam pra-pemrosesan data medis, yang memungkinkan peneliti untuk fokus pada informasi yang paling signifikan dan mengabaikan noise atau informasi yang tidak relevan yang dapat mengaburkan pola yang sebenarnya ada dalam data. Dengan menerapkan metode pemilihan fitur yang tepat, peneliti dan praktisi dapat meningkatkan kualitas model prediktif dan memastikan bahwa hanya informasi yang paling relevan yang digunakan untuk membuat keputusan yang berdampak pada pasien (Leto et al., 2023).

Teknik data mining yang digunakan pada penelitian ini adalah teknik *feature selection*. Seleksi fitur adalah salah satu teknik data mining yang umum digunakan pada tahapan *pre-processing*. Teknik ini digunakan untuk mengurangi kompleksitas atribut yang akan dikelola pada *processing* dan analisis. Teknik ini dilakukan untuk mengetahui subset fitur yang paling signifikan dari dataset *Primary Tumor*. Pemilihan fitur sering digunakan untuk pengurangan dimensi model. Pemilihan fitur membantu mengurangi fitur domain, menghilangkan fitur yang berlebihan. Dengan cara ini akan membantu mempercepat proses pembelajaran atau pemodelan (Han et al., 2012).

Metode *backward* dan metode *stepwise* adalah dua pendekatan yang sering digunakan dalam pemilihan fitur. Metode backward melibatkan penghapusan iteratif fitur-fitur yang dianggap kurang relevan satu per satu hingga hanya fitur-fitur yang paling penting yang tersisa. Sementara metode stepwise melibatkan proses iteratif di mana fitur-fitur ditambahkan atau dihapus berdasarkan perbandingan kriteria tertentu, seperti kriteria Akaike atau kriteria informasi bayes.

Ketika datang ke analisis data medis, seperti klasifikasi jenis tumor primer, *accuracy* adalah metrik kinerja yang sangat kritis. Kemampuan model untuk dengan benar mengklasifikasikan jenis tumor primer memiliki dampak langsung pada diagnosis dan pengobatan pasien. Oleh karena itu, perbandingan antara metode pemilihan fitur dalam hal accuracy adalah penting untuk memahami mana yang lebih efektif dalam menghasilkan model yang akurat dan dapat diandalkan.

Pemilihan atribut yang tepat adalah kunci untuk membangun model klasifikasi yang efektif, dan dalam penelitian ini, kami memfokuskan pada pemilihan atribut pada dataset *Primary Tumor*, yang berisi informasi tentang tumor primer. Kami akan membandingkan dua *Feature Selection*, yaitu *Backward Method* dan *Stepwise Method*, dalam konteks pemodelan klasifikasi. *Backward method* menghapus atribut dengan kontribusi terendah, sementara metode *Stepwise method* menambah atau mengurangi atribut berdasarkan pengaruhnya pada klasifikasi. Hasilnya akan membantu pemilih model dan peneliti dalam memilih metode yang sesuai untuk studi mereka dan merangsang penelitian lanjutan dalam seleksi atribut pada data medis serupa.

2. METODE PENELITIAN

Pada penelitian ini dilakukan beberapa tahapan penelitian yaitu pengumpulan data, *pre-processing*, dan perbandingan *feature selection* menggunakan model klasifikasi regresi logistik.

2.1 Sumber Data

Dataset yang digunakan dalam preprocessing ini yaitu *Primary Tumor* yang diambil dari *UCI Machine Learning Repository*. UCI (*University of California, Irvine*) *Machine Learning Repository* adalah salah satu sumber terkemuka untuk dataset yang digunakan dalam penelitian ilmu data dan pembelajaran mesin. Dataset ini mungkin telah dikumpulkan dari berbagai sumber, termasuk lembaga medis, penelitian ilmiah, atau institusi kesehatan. *UCI Machine Learning Repository* bertujuan untuk menyediakan akses terbuka kepada berbagai dataset yang digunakan oleh peneliti dan praktisi di berbagai bidang, termasuk ilmu data, pembelajaran mesin, dan analisis statistik

2.2 Dataset Primary Tumor

Dataset yang digunakan pada penelitian ini yaitu Primary Tumor. Dataset ini memiliki 339 data dengan 18 Atribut (*class, age, sex, histologic-type, degree of diffe, bone, bone marrow, lung, pleura, peritoneum, liver, brain, skin, neck, supraclavicular, axillar, mediastrium, abdominal*). Adapun variabel dataset yang di gunakan pada penelitian dapat di lihat pada tabel 1.

Tabel 1. Variabel dataset Primary Tumor

No	Atribut	Keterangan	No	Atribut	Keterangan
1.	Class	Bagian tubuh tempat terjadinya tumor	10.	Peritoneum	Mengindikasikan apakah terdapat penyebaran tumor ke peritoneum (Ya/Tidak)
2.	Age	Usia pasien pada saat diagnosis tumor.	11.	Liver	Mengindikasikan apakah terdapat penyebaran tumor ke hati (Ya/Tidak)
3.	Sex	Jenis kelamin pada pasien (Pria/wanita)	12.	Brain	Mengindikasikan apakah terdapat penyebaran tumor ke otak (Ya/Tidak)
4.	Histologic -type	Tipe-tipe ini mencakup epidermoid, adeno dan anaplastic	13.	Skin	Mengindikasikan apakah terdapat penyebaran tumor ke kulit (Ya/Tidak)
5.	Degree of diffe	Derajat diferensiasi dapat digambarkan sebagai baik, cukup, atau buruk	14.	Neck	Mengindikan apakah terdapat penyebaran tumor ke leher (Ya/Tidak)
6.	Bone	Mengindikasikan apakah terdapat penyebaran tumor ke tulang (Ya/Tidak)	15.	Supraclavicular	Mengindikasikan apakah terdapat penyebaran tumor ke supraclavicular (Ya/Tidak)
7.	Bone-marrow	Mengindikasikan apakah terdapat penyebaran tumor ke sumsum tulang (Ya/Tidak)	16.	Axillar	Mengindikasikan apakah terdapat penyebaran tumor ke ketiak (Ya/Tidak)
8.	Lung	Mengindikasikan apakah terdapat penyebaran tumor ke paru-paru (Ya/Tidak)	17.	Mediastrium	Mengindikasikan apakah terdapat penyebaran tumor ke mediastinum (Ya/Tidak)
9.	Pleura	Mengindikasikan apakah terdapat penyebaran tumor ke pleura (Ya/Tidak)	18.	Abdominal	Mengindikasikan apakah terdapat penyebaran tumor ke rongga perut, (Ya/Tidak)

2.3 *Pre-Processing*

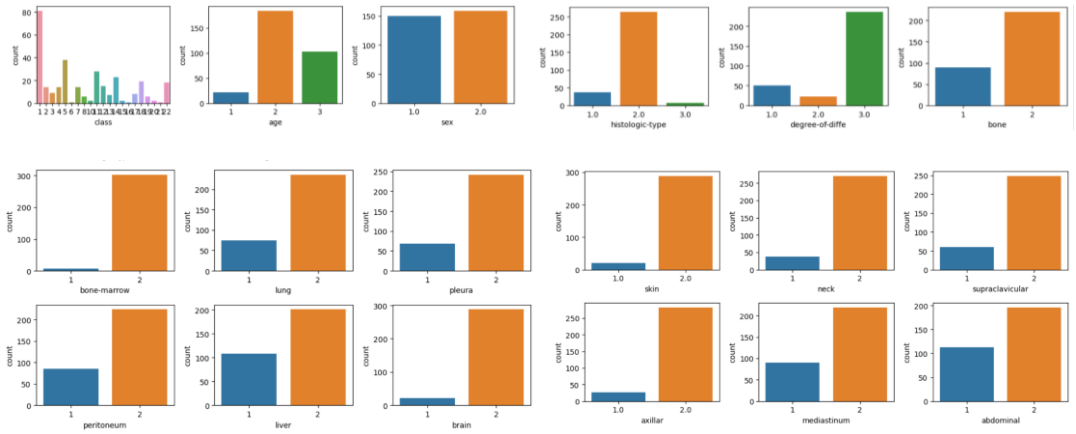
Preprocessing adalah suatu teknik untuk membuat data menjadi lebih mudah diproses atau digunakan dalam data mining. Menurut Han, Kamber, dan Pei (2011), Tujuan dari *pre-processing* ini untuk membuat kualitas data yang baik, termasuk kelengkapan, konsistensi, ketepatan waktu *dan* meningkatkan hasil akurasi. Adapun beberapa tahapan preprocessing yang dilakukan pada dataset *Primary Tumor*:

2.3.1 Data *Cleaning*

Data cleansing adalah proses memodifikasi atau menghapus data yang dianggap tidak akurat, duplikat, tidak lengkap, salah format, maupun rusak dalam kumpulan data yang dimiliki.

1. *Duplicate Data*
   Duplikat data merupakan sebuah data yang mirip dan sengaja disamakan dengan data aslinya. Pada dataset "Primary Tumor" terdapat 30 duplikat data dengan jumlah data sebanyak 339 yang kemudian setelah dihapus menjadi 309 data.
2. *Missing Value*
   Missing Value merupakan hilangnya beberapa data yang telah diperoleh. Pada dataset ini terdapat 5 atribut yang memiliki missing value yaitu *sex, degree-of-diffe, histologic type, skin, dan axillar*. Salah satu cara yang dapat dilakukan untuk menangani missing data adalah dengan mengisi missing data dengan nilai-nilai yang mungkin berdasarkan informasi yang tersedia pada data atau dikenal dengan imputasi. Pada kasus ini *missing value* diatasi dengan mengganti "?" dengan "NA" lalu mengisinya dengan data yang paling sering muncul (modus) di kolom tersebut. Pemilihan metode ini berdasarkan dengan tipe data, yaitu kategorikal.
3. *Outlier*
   Outlier merupakan nilai yang ekstrem atau tidak biasa, dan dapat memiliki dampak signifikan pada hasil analisis statistik jika tidak dikelola dengan baik. Outlier dapat muncul dalam berbagai jenis data, termasuk data numerik dan data kategorikal. Outlier dapat muncul karena beberapa alasan, termasuk kesalahan pengukuran, variasi alamiah dalam data, atau bahkan sebagai hasil dari peristiwa yang tidak biasa atau langka dalam data. Berikut merupakan hasil outlier dari dataset primary tumor, dapat dilihat pada gambar 1.

Gambar 1. Outlier pada dataset Primary Tumor



2.3.2 Transformasi Data

Transformasi Data adalah upaya yang dilakukan dengan tujuan utama untuk mengubah skala pengukuran data asli menjadi bentuk lain sehingga data dapat memenuhi asumsi-asumsi yang mendasari analisis ragam. Terdapat 2 cara untuk melakukan transformasi data yaitu *MinMax Normalization* dan *Z-Score Normalization*. Pada penelitian ini menggunakan *Minmax Normalization* untuk transformasi data. *Minmax Normalization* adalah teknik normalisasi yang digunakan dalam data preprocessing untuk mengubah rentang nilai data menjadi rentang yang telah ditentukan, biasanya dari 0 hingga 1. Tujuan dari *Min-Max Normalization* adalah untuk memastikan bahwa semua atribut (fitur) dalam dataset memiliki perbandingan skala yang serupa. Min-max normalization dapat dihitung menggunakan rumus berikut:

$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Berikut merupakan data *Min-max normalization* pada dataset Primary dapat dilihat pada tabel 2.

Tabel 2. Data MinMax Normalization

	Age	Sex	Histologic-type	Degree of diffe	...	Abdominal
Count	309.000000	309.000000	309.000000	309.000000	...	309.000000
Mean	0.631068	0.514563	0.453074	0.800971	...	0.634304
Std	0.290218	0.500599	0.185248	0.375853	...	0.482406
Min	0.000000	0.000000	0.000000	0.000000	...	0.000000
25%	0.500000	0.000000	0.500000	1.000000	...	0.000000
50%	0.500000	1.000000	0.500000	1.000000	...	1.000000
75%	1.000000	1.000000	0.500000	1.000000	...	1.000000
max	1.000000	1.000000	1.000000	1.000000	...	1.000000

2.4 *Feature Selection*

Feature selection adalah suatu proses menghapus features yang berlebihan dan tidak relevan dari dataset yang sebenarnya. Sehingga waktu yang digunakan mengeksekusi dari pengklasifikasi yang memproses data berkurang, dan dapat meningkatkan akurasi juga karena features yang tidak relevan dapat memperburuk data mempengaruhi akurasi klasifikasi secara negatif (Doraisami & Golzari, 2008). Dengan *feature selection* dapat meningkatkan pemahaman dan biaya penanganan data menjadi lebih kecil (Arauzo et. al., 2011). *Feature Selection* yang digunakan dalam penelitian ini terdiri dari 2 metode, yaitu *backward method*, dan *stepwise method*.

2.4.1 *Backward Method*

Metode backward merupakan metode yang memiliki fungsi untuk mengoptimalkan kinerja suatu model dengan sistem kinerja mundur. Pemilihan dilakukan dengan cara memilih variabel ke depan yaitu dengan menguji semua variabel kemudian menghapus variabel-variabel yang dianggap tidak relevan (Nugroho, 2020). Berikut Langkah-langkah metode *Backward Elimination*

a. Membuat model degan meregredikan variabel reson Y dengan semua variabel prekdiktor.
b. Mengeluarkan satu-saru vaiabel *predictor* dengan melakukan pengujian terhadap parameternya dengan partial Ftest. Niali Fparsial terkecil dibandingkan dengan Ftabel :
    1. Jika Fparsial < Ftabel, maka X yang bersangkutan dikeluakan dari model dan dilanjutkan dengan pembuatan model baru tanpa variable tersebut.
    2. Jika Fparsial >Ftabel, maka proses dihentikan artinya tidak ada variabel yang perlu dikeluarkan dan persamman terakhir yang dipilih.

2.4.2 *Stepwise Method*

Metode stepwise adalah suatu pendekatan dalam analisis statistik yang digunakan untuk memilih model regresi yang paling sesuai dengan data. Pendekatan ini melibatkan penambahan atau penghapusan variabel secara bertahap dari model regresi untuk menentukan variabel mana yang paling signifikan dalam memprediksi variabel target.

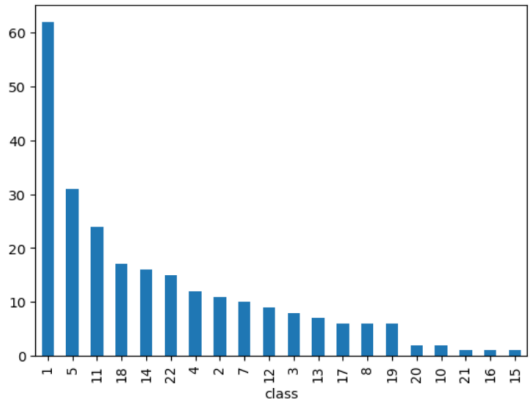
2.5 *Split Data*

Data splitting atau pemisahan data adalah metode membagi data menjadi dua bagian atau lebih yang membentuk subhimpunan data. Umumnya, data splitting memisahkan dua bagian, bagian pertama digunakan untuk mengevaluasi atau uji data dan data lainnya digunakan untuk melatih model. Pada dataset Primary Tumor, dilakukan pembagian dataset dengan proporsi tertentu untuk training dan testing. Proporsi yang digunakan adalah 20% untuk testing set dan 80% untuk training set. Dengan demikian, terdapat dua subset utama: data training yang terdiri dari 247 data, dan data testing yang memiliki 62 data.

2.6 *Imbalance Data*

Imbalance data merupakan kondisi dimana suatu kelompok kelas memiliki jumlah data yang jauh berbeda dibandingkan dengan kelas lainnya. Kelas yang memiliki jumlah data lebih banyak disebut dengan *majority class* dan kelas yang mempunyai jumlah data lebih sedikit disebut dengan *minority class* (Barro dkk., 2013). Berikut merupakan imbalance data dari dataset Primary tumor dapat dilihat pada gambar 1.

Gambar 2. Barchart Imbalance data Primary Tumor



Grafik di atas menunjukkan adanya ketidakseimbangan (*imbalance*) dalam data. Dapat dilihat bahwa nilai dengan jumlah paling besar berada pada class 1 yaitu 62 data dan nilai dengan jumlah paling kecil berada pada *class* 15, *class* 16, *class* 21 yaitu 1 data. Ketidakseimbangan data dapat menjadi masalah serius dalam pembelajaran mesin, karena model yang dilatih dengan data yang tidak seimbang cenderung menjadi bias terhadap kelas mayoritas, sementara kelas minoritas dapat diabaikan. Untuk mengatasi ketidakseimbangan data, salah satu teknik yang digunakan adalah teknik *Random Oversampling* (ROS). Teknik *Random Oversampling* (ROS) bertujuan untuk mengatasi masalah ketidakseimbangan dengan meningkatkan jumlah data pada kelas minoritas. Ini dilakukan dengan cara menggandakan atau mengulangi data yang ada dalam kelas minoritas secara acak, sehingga jumlahnya menjadi lebih seimbang dengan kelas mayoritas. Dengan menggunakan ROS, model pembelajaran mesin memiliki lebih banyak contoh untuk memahami dan mempelajari pola dalam kelas minoritas, sehingga meningkatkan kemampuan model dalam memprediksi data pada kelas tersebut.

2.7 Membandingkan Feature Selection

Setelah menangani ketidakseimbangan data dengan teknik oversampling, langkah selanjutnya adalah melakukan perbandingan antara dua metode *feature selection*, yaitu metode *backward* dan metode *stepwise*. Setelah mengaplikasikan kedua metode seleksi fitur ini, dilanjutkan dengan pelatihan model regresi logistik. Hasil dari pelatihan model tersebut kemudian dievaluasi dengan melihat accuracy dari masing-masing metode seleksi fitur. Metode seleksi fitur adalah langkah penting dalam pengembangan model, karena dapat membantu memilih subset fitur terbaik yang memberikan kontribusi paling signifikan dalam memprediksi target. Perbandingan accuracy antara metode *backward* dan metode *stepwise* dapat membantu menilai kualitas dan efektivitas masing-masing metode dalam meningkatkan kinerja model.

3. HASIL DAN PEMBAHASAN

Berdasarkan tujuan dari penelitian ini yaitu membandingkan kinerja dua *Feature Selection*, antara *Backward Method* (SBS) dan *Stepwise Method* (SFFS) untuk mengetahui nilai *accuracy*, pada dataset Primary Tumor. Hasil kedua metode tersebut telah memilih subset variabel yang baik dan signifikan. Namun, terdapat juga beberapa perbedaan, seperti:

3.1 Jumlah Variabel yang Dipilih:

Dalam backward, metode tersebut terdapat 13 variabel terbaik yaitu: *age, sex, histologic type, degree of diffe, bone, lung, liver, brain, neck, supraclavicular, axillar, mediastinum,* dan *abdominal*. Sedangkan dalam stepwise, metode tersebut terdapat 13 variabel terbaik yaitu: *age, sex, histologic type, degree of diffe, bone, lung, peritoneum, skin, neck, supraclavicular, axillar, mediastinum* dan *abdominal*. Dalam kasus ini, kedua metode menghasilkan daftar 13 variabel terbaik yang dianggap berperan penting. Berdasarkan *feature selection* yang meliputi backward method dan stepwise method terdapat 11 atribut yang tetap atau sering muncul yang mempunyai korelasi yang besar dan dianggap sebagai atribut yang penting terhadap variabel Y ('class') yaitu *age, sex, histologic type, degree of diffe, bone, lung, neck, supraclavicular, axiilar, mediastinum, dan abdominal.*

3.2 Pemilihan Model Klasifikasi

Dalam penelitian ini, model klasifikasi yang digunakan adalah Model Regresi Logistik. Menggunakan model klasifikasi regresi logistik merupakan salah satu cara untuk membandingkan metode seleksi variabel *backward method* dan *stepwise method*. Data train yang digunakan dalam model ini adalah dataset yang digunakan untuk melatih model regresi logistik. Ketika membandingkan metode seleksi variabel seperti *backward method* dan *stepwise method*, data train akan digunakan untuk melatih model regresi logistik dalam dua skenario yang berbeda. Pertama, dengan melatih model regresi logistik dengan menggunakan variabel-variabel yang dipilih dengan metode *backward*. Kedua, dengan melatih model yang serupa dengan menggunakan variabel-variabel yang dipilih dengan metode *stepwise*. Kemudian, hasil performa kedua model ini akan dibandingkan untuk melihat mana yang lebih baik dalam mengklasifikasikan data test yang belum dilihat sebelumnya. Hasil perbandingan performa model regresi logistik yang menggunakan kedua metode seleksi variabel ini akan membantu dalam menentukan metode mana yang lebih efektif dan akurat dalam mengidentifikasi variabel yang paling relevan dalam hubungannya dengan variabel "class" pada dataset yang diberikan.

3.3 Hasil Membandingkan Feature Selection

Dalam perbandingan feature selection antara backward dan stepwise menggunakan regresi logistik dengan melihat accuracy menggunakan metrik cross-validation score, dilakukan evaluasi kualitas model yang telah dibangun. Cross-validation adalah metode yang digunakan untuk mengukur sejauh mana model statistik atau prediktif mampu menggeneralisasi hasilnya ke data yang belum pernah dilihat sebelumnya. Metrik yang sering digunakan dalam cross-validation adalah akurasi, yang mengukur sejauh mana model dapat memprediksi dengan benar berdasarkan data yang tidak digunakan dalam pelatihan. Accuracy cross-validation score mencerminkan sejauh mana model mampu memprediksi dengan benar pada data uji yang tidak digunakan dalam pelatihan. Berikut merupakan hasil accuracy dari kedua metode dapat dilihat pada tabel 3.

Tabel 3. Hasil Accuracy Backward Method dan Stepwise Method

Model	*Accuracy*
Backward Method	66.774 (0.021)
Stepwise Method	65.806 (0.008)

Untuk menuliskan persamaan atau rumus-rumus statistika, disarankan untuk menggunakan Microsoft Equation Editor untuk menulis setiap rumus atau persamaan yang muncul dalam teks. Pada tabel diatas dapat dilihat bahwa, metode backward menghasilkan accuracy sebesar 66.774 (0.021), sedangkan metode stepwise menghasilkan accuracy sebesar 65.806 (0.008). Dari hasil ini, dapat disimpulkan bahwa metode backward menghasilkan model dengan accuracy yang sedikit lebih tinggi dibandingkan dengan metode stepwise

Namun, perbedaan dalam accuracy antara kedua metode tersebut tidak terlalu signifikan, dan tingkat ketepatan prediksi pada kedua model cukup mendekati satu sama lain. Selain itu, deviasi standar yang relatif rendah menunjukkan konsistensi performa model, yang merupakan indikator positif. Dengan demikian, membandingkan kedua metode dengan akurasi cross-validation score memberikan wawasan tentang sejauh mana metode seleksi variabel berkontribusi terhadap hasil akhir model regresi logistik, membantu kita dalam memahami kualitas dan kehandalan model yang telah dibangun.

4. SIMPULAN DAN SARAN

Berdasarkan pembagian data testing dan training dengan rasio 20:80 serta perbandingan kinerja dua metode Feature Selection antara Backward Method (SBS) dan Stepwise Method (SFFS) menggunakan model klasifikasi regresi logistik, diperoleh hasil accuracy. Accuracy yang dihasilkan adalah 66.774 (0.021) untuk metode Backward (SBS) dan 65.806 (0.008) untuk metode Stepwise (SFFS). Hal ini menunjukkan bahwa metode backward menghasilkan model dengan accuracy yang sedikit lebih tinggi dan mampu memberikan prediksi yang lebih baik dibandingkan dengan metode stepwise.

Saran untuk penelitian selanjutnya adalah dapat dillakukan mencari nilai *accuracy* dengan membandingkan metode lain dari *feature selection* seperti *forward method* dan *filter method.* Dapat juga menggunakan model selain regresi logistik seperti Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbour, dan Artificial Neural Network. Selain itu, dapat membandingkan sampling dengan membandingkan hold out dan k-fold, atau membandingkan stratifikasi s holdout dan s k-fold. Ketiga konsep tersebut dapat dilakukan untuk modelling klasifikasi pada dataset

DAFTAR PUSTAKA

Arie, Gunawan (2023) *Pengantar Sistem Informasi Kesehatan.* PT. Literasi Nusantara Abadi Grup.

Arauzo-Azofra, J. L. Aznarte, and J. M. Benítez, "Empirical study of feature selection methods based on individual feature evaluation for classification problems," Expert Systems with Applications, Vol, 38. 8170-8177, 2011.

Barro, R. A., Sulvianti, I. D., dan Afendi, F. M. 2013. *Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu*. Journal of Statistics. 1(1).

Han, J., Kamber, M., and Pei, J., 2011. *Data Mining Concepts and Techniques*. (3rd ed.). USA: Morgan Kaufmann.

Han, J., Kamber, M., and Pei, J., 2012, *Data Mining*: Concepts and Techniques (Waltham, MA: Elsevier).

Leto, C., Sujana, D., Windyasari, V. S., Mahmudin, & Muhammad, R. (2023). KONSEP DATA MINING DAN PENERAPAN.

Nugroho, W. (2020). Optimasi Metode K-Nearest Neighbours dengan Backward Elimination Menggunakan Dataset Software Effort Estimation Bianglala Informatika. *Bianglala Informatika*, *8*(2), 129–133.

S. Doraisami, dan S. Golzari, "*A Study on Feature selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music, Content-Based Retrieval, Categorization and Similarity*" 2008.

Zwitter, M., dan M., Soklic. (1988). Primary Tumor. https://archive.ics.uci.edu/dataset/83/primary+tumor.