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

PERBANDINGAN METODE FEATURE SELECTION *FILTER METHOD* DAN *BACKWARD METHOD* PADA DATASET *DERMATOLOGY*

NISWA NILHAYA M.[1], ROSY MARGRETH LATUNUSA[2], CAECILIA BINTANG GIRIK ALLO[3]

[1)]Program Studi Statistika Fakultas MIPA Universitas Cenderawasih Jayapura, Indonesia
[2)] Program Studi Statistika Fakultas MIPA Universitas Cenderawasih Jayapura, Indonesia
[3)] Program Studi Statistika Fakultas MIPA Universitas Cenderawasih Jayapura, Indonesia
e-mail: niswanilhaya@gmail.com[1],rosylatunusa12@gmail.com[2],bintanggirikallo@gmail.com[3]

ABSTRAK

Dermatologi merupakan cabang ilmu kedokteran yang mempelajari penyakit kulit yang memiliki gejala klinis serupa, seperti kemerahan dan bersisik, namun dengan sedikit perbedaan. Beberapa penyakit yang termasuk dalam kelompok ini antara lain *psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, dan pityriasis rubra pilaris*. Salah satu metode komputasi yang dapat digunakan dalam pengolahan data adalah data *mining*. Dalam data *mining*, data *pre-processing* merupakan tahapan yang sangat penting. Salah satu teknik data *pre-processing* yang sering digunakan untuk mengetahui atribut yang paling berpengaruh pada sebuah dataset adalah *feature selection*. Metode *feature selection* dapat membantu dalam perbandingan fitur yang paling sesuai dengan melihat nilai akurasi dari tiap metode *feature selection* yang digunakan. Dataset *Dermatology* yang digunakan dalam penelitian ini adalah data yang diunduh dari situs *repository UCI Machine Learning*. Dari hasil pengujian perbandingan metode *feature selection* yaitu *filter method* dan *backward method* diperoleh metode *feature selection* terbaik yaitu pada *backward method* dengan nilai akurasi tertinggi yaitu 98.286

Kata Kunci: dermatology, feature selection, filter method, backward method

1. PENDAHULUAN

Dermatologi merupakan cabang ilmu kedokteran yang mempelajari penyakit kulit. Salah satu tantangan dalam dermatologi adalah mendiagnosis psoriasis eritematosa (ESD), yang memiliki gejala klinis serupa, seperti kemerahan dan bersisik, namun dengan sedikit perbedaan. Beberapa penyakit yang termasuk dalam kelompok ini antara lain psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, dan pityriasis rubra pilaris. (UCI Machine Learning, 2023).

Dalam pengolahan data, salah satu metode komputasi yang dapat digunakan adalah data mining. Perkembangan alat prediktif dapat membantu dokter mendiagnosis pasien dengan lebih cepat dan efektif. Data mining telah terbukti menjadi pendekatan yang kuat dan efektif yang menyediakan proses untuk menemukan pola dalam kumpulan data yang besar (Jaree Thongkam, G. X, 2008).

Dalam data *mining*, data *pre-processing* merupakan tahapan yang sangat penting. *Pre-processing* berguna untuk mempersiapkan data sehingga teknik data mining yang diterapkan menghasilkan pola yang berkualitas dan akurat (Jasmina Nalic, A. S., 2018). Salah satu teknik data *pre-processing* yang sering digunakan untuk mengetahui atribut yang paling berpengaruh pada sebuah dataset adalah feature selection. Teknik ini digunakan untuk mengurangi kompleksitas atribut yang akan diolah dan dianalisis (Adnyana, 2019). Dengan menggunakan dataset ini, metode feature selection dapat membantu dalam perbandingan fitur yang paling sesuai dengan melihat nilai akurasi dari tiap metode *feature selection* yang digunakan.

2. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini terdiri dari data *preparation*, data *pre-processing*, dan data *processing*.

2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah data Dermatology yang diunduh dari *repository UCI Machine learning*. Dataset penyakit kulit ini berisi gejala-gejala yang terjadi pada penyakit kulit. Dataset ini terdiri dari 35 variabel, dengan 34 variabel X dan 1 variabel Y dan total 366 baris.

2.2 Pre-pocessing

Pada tahap *pre-processing*, pembersihan data perlu dilakukan untuk membersihkan nilai-nilai yang hilang pada data. Pada tahap ini dilakukan pengecekan apakah terdapat *missing value* pada dataset. kemudian apabila telah diketahui nilai *missing value* pada suatu variabel, maka pada data tersebut variabel yang diketahui memiliki *missing value* tersebut tidak dihapus melainkan tetap digunakan dengan cara mengganti data pada variabel tersebut dengan nilai yang hilang tersebut dengan menggunakan mean maupun median.

Correlation Attribute Evaluation merupakan metode pemilihan fitur yang menggunakan metode pencarian peringkat. *Correlation Attribute Evaluation* memperhatikan kelas sasaran. Korelasi setiap atribut dengan kelas sasaran diukur menggunakan metode korelasi *Pearson*. Setiap nilai berperan sebagai indikator dengan mempertimbangkan atribut nominal dalam basis nilai.

Langkah pertama adalah menghitung korelasi pada kumpulan data menggunakan persamaan (1). Korelasi Pearson disebut parameter populasi. Metode ini digunakan jika kedua variabel penelitian berdistribusi normal. Nilai koefisien dipengaruhi oleh nilai ekstrim yang dapat menambah atau mengurangi kekuatan hubungan. Oleh karena itu, tidak tepat jika salah satu atau kedua variabel tidak berdistribusi normal.

$$\rho i = \frac{cov(X_i,Y)}{\sigma(X_i)\sigma_Y} \quad (1)$$

Yang paling berdampak pada target, koefisien korelasi Pearson antara Fitur dan target dituliskan dalam persamaan (1), di mana covariance merupakan standar deviasi; memiliki kisaran antara 0 dan 1.

Feature selection merupakan teknik penting dan sering digunakan dalam tahap pre-processing. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan nilai kelas target. Fitur yang diabaikan seringkali merupakan fitur yang tidak relevan dan data yang berlebih. Tujuan utama feature selection adalah memilih fitur terbaik dari kumpulan data fitur. Secara umum, metode feature selection dapat dibagi menjadi tiga kelompok, yaitu filter, wrapper, dan embedded selector.

Metode filter mengevaluasi setiap fitur secara independen dari pengklasifikasi, yang kemudian memberi peringkat pada fitur setelah evaluasi dan memilih fitur unggulan. Metode filter menerapkan ukuran statistik untuk memberikan skor pada setiap fitur. Metode pfilter menggunakan kriteria evaluasi meliputi jarak, keinformatifan, keandalan, dan konsistensi. Metode penyaringan menggunakan kriteria utama teknik perangkingan dan menggunakan urutan perangkingan untuk pemilihan variabel.

Metode wrapper membutuhkan algoritma Machine Learning dan menggunakan kinerjanya sebagai kriteria evaluasi. Metode ini mencari fitur yang paling relevan dengan algoritma Machine Learning dan bertujuan untuk meningkatkan kinerja algoritma. Untuk mengevaluasi fitur, akurasi prediksi digunakan dalam tugas klasifikasi. Metode wrapper untuk pemilihan fitur dapat dibagi menjadi tiga kategori: Forward selection, Backward elimination dan Recursive Feature elimination.

Forward selection merupakan metode pemilihan berulang dimulai dengan fitur kosong dalam model. Dengan setiap iterasi, akan ditambahkan fitur yang memiliki dampak paling signifikan terhadap peningkatan model yang dimiliki. Selanjutnya, penambahan variabel baru tidak meningkatkan performa model. Backward Elimination merupakan kebalikan dengan metode forward selection, pada metode ini model memuat semua fitur. Kemudian pada setiap iterasi, fitur yang tidak meningkatkan performa model secara signifikan akan dihapus. Proses ini diulangi hingga model berisi fitur-fitur ideal, yang ditunjukkan dengan tidak ditemukannya perubahan ketika fitur-fitur tersebut dihapus. Recursive Feature elimination adalah metode optimasi algoritma greedy yang bertujuan untuk menemukan subset fitur berkinerja terbaik.

Pada setiap iterasi, metode ini membangun model yang dimulai dari fitur paling kiri sampai semua fitur selesai dijelajahi. Metode ini mengabaikan fitur berkinerja terbaik atau terburuk di setiap iterasi. Sebaliknya metode ini memberi peringkat fitur berdasarkan urutan eliminasinya.

2.3 Processing

Proses pengolahan data mining pada tahap imbalance data melibatkan langkah-langkah khusus untuk menangani ketidakseimbangan kelas dalam dataset. Beberapa teknik yang dapat digunakan untuk mengatasi masalah ini seperti Oversampling. Teknik oversampling melibatkan peningkatan jumlah sampel dari kelas minoritas dalam dataset. Metode oversampling yang umum digunakan antara lain Random Oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), dan Majority Weighted Minority Oversampling Technique (MWMOTE). Selain oversampling terdapat juga teknik undersampling. Teknik undersampling melibatkan pengurangan jumlah sampel dari kelas mayoritas dalam kumpulan data. Namun, pengambilan sampel yang terlalu rendah dapat menyebabkan kelompok mayoritas kehilangan informasi berharga, sehingga hal ini tidak selalu merupakan pilihan yang baik.

Penentuan model berdasarkan nilai akurasi dilakukan terlebih dahulu Pemilihan metode klasifikasi yaitu regresi logistik. Pada algoritma regresi logistik, variabel independen didefinisikan sebagai persamaan regresi linear

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2)$$

Algoritma regresi logistik memprediksi probabilitas keanggotaan variabel independen dalam suatu kelas menggunakan model fungsi logit transform atau invers logit dengan persamaan.

Dalam tahap penentuan model berdasarkan nilai akurasi, semakin tinggi nilai akurasi yang diperoleh, maka semakin baik model yang dihasilkan dari proses data mining tersebut. Oleh karena itu, pemilihan metode klasifikasi dan metrik evaluasi yang tepat sangat penting untuk memastikan model yang dihasilkan akurat dan dapat diandalkan.

3. HASIL DAN PEMBAHASAN

Tahapan analisis dimulai dengan tahap penyiapan data, preprocessing, pengolahan, dan klasifikasi. Data yang telah disiapkan kemudian dibersihkan dengan memeriksa duplikat data. Pada dataset *Dermatology* yang digunakan tidak terdapat data duplikat. Data kemudian diperiksa apakah terdapat *missing value*, diperoleh salah satu variabel terdapat *missing value*, yaitu variabel *Age (linier)* dengan total delapan (8) *missing value*. Variabel *Age(linier)* berisi data yang bersifat numerik, sehingga untuk mengisi *missing value* pada variabel ini gunakan nilai median. Berikutnya dilakukan deteksi *outlier*, hanya variabel *Age (linear)* yang merupakan data numerik, sehingga deteksi *outlier* hanya pada variabel *Age (linear)* dan variable *Age(linear)* tidak terdapat *outlier*. Selanjutnya, untuk menyeragamkan data dengan skala [0 , 1] dilakukan standarisasi data dengan *Min-Max normalization*. Berikutnya dilakukan, *feature selection* untuk mengurangi variabel yang tidak relevan dan tidak memberikan pengaruh signifikan dalam meningkatkan kinerja model. Metode yang digunakan dalam f*eature selection* antara lain *filter method, forward method, backward method* dan *stepwise*. Untuk perbandingan metode yang akan digunakan yaitu *filter method* dan *backward* dengan model klasifikasi *Logistic Regression*.

Berikut variabel-variabel yang tersisa pada metode *filter method* dan *backward*.

Tabel 1. Filter Method

Filter Method					
erythema	knee and elbow involvement	fibrosis of the papillary dermis	elongation of the rete ridges	disappearance of the granular layer	Age (linear)
scaling	scalp involvement	exocytosis	thinning of the suprapapillary epidermis	spongiosis	Class

definite borders	family history	parakeratosis	spongiform pustule	follicular horn plug	
follicular papules	PNL infiltrate	clubbing of the rete ridges	munro microabcess	perifollicular parakeratosis	

Tabel 2. Backward

Backward			
erythema	koebner phenomenon	PNL infiltrate	spongiosis
scaling	follicular papules	fibrosis of the papillary dermis	
definite borders	oral mucosal involvement	parakeratosis	
itching	melanin incontinence	disappearance of the granular layer	

Dataset yang telah melalui proses *pre-processing* selanjutnya masuk pada tahap *processing* yaitu split data, data akan diuji dengan membagi menjadi dibagi menjadi dua bagian yaitu data train dan data test. Berikutnya dilakukan pengecekan terhadap data apakah data berada pada kondisi *balance* (seimbang) atau data tidak *balance*. Dilakukan pegecekan diperoleh bahwa data berada pada kondisi tidak *balance* sehingga langkah selanjutnya adalah membuat data menjadi *balance* dengan menggunakan metode SMOTE. Seletah data seimbang dilanjutkan pada langkah model klasifikasi yaitu digunakan model klasifikasi Logistic Regression. Diperoleh nilai akurasi regresi logistic pada metode *filter* sebesar 91.074 dan nilai akurasi uregresi logistic pada metode *backward* sebesar 98.286. Metode terbaik adalah metode dengan nilai akurasi tertinggi. Dengan demikian diperoleh metode terbaik dengan nilai akurasi tertinggi metode *backward*.

4. SIMPULAN DAN SARAN

Pada dataset Dermatology dilakukan perbandingan metode feature selection. Dari empat metode yang terdapat pada feature selection dipilih dua metode sebagai perbandingan yaitu filter method dan backward. Setelah dilakukan perhitungan diperoleh nilai akurasi pada filter method sebesar 91.074 dan nilai akurasi pada backward sebesar 98.286. Dari nilai akurasi akan ditentukan metode terbaik. Metode terbaik dipilih berdasarkan nilai akurasi tertinggi, sehingga metode terbaik pada dataset Dermatology yaitu backward dengan nilai akurasi tertinggi yaitu 98.286.

DAFTAR PUSTAKA

Adnyana, I. M. (2019). Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa. *JURNAL SISTEM DAN INFORMATIKA*.

Annisa Nurul Puteri, A. A. (2021). Feature Selection Correlation-Based pada Prediksi Nasabah Bank Telemarketing untuk Deposito. *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*.

Anurag Kumar Vermaa, S. P. (2019). Comparison of skin disease prediction by feature selection using ensemble.

Jaree Thongkam, G. X. (2008). Breast Cancer Survivabilityvia AdaBoost Algorithms. *Second Australian Workshop on Health Data and Knowledge Management*.

Jasmina Nalic, A. S. (2018). Importance of Data Pre-Processing in Credit Scoring Models Based on Data Mining Approaches. *Computer Science*.

Khafid Akbar, M. H. (2020). Data Balancing untuk Mengatasi Imbalance Dataset pada Prediksi Produksi Padi. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*.

Prajarini, D. (2016). Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit. *Informatics Journal*.

Reza Dwi Fitriani, H. Y. (2021). PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES. *JURNAL GAUSSIAN*.

UCI. (2023, September). Retrieved from UCI MACHINE LEARNING: https://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/dermatology.data

WICKLY GUSTHVI, A. A. (2023). Perbandingan Metode Klasifikasi Decission Tree, Naive Bayes, K-Nearest-Neighbor, dan Logistic Regression pada Dataset Phishing. *CENDERAWASIH Journal of Statistics and Data Science*.