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

ANALISIS PERBANDINGAN METODE FEATURE SELECTION BACKWARD METHOD DAN STEPWISE METHOD

NATASYA I. PARENDEN[1], NURFADILLAH[2], MARIA F. BAREK BUNGA[3], MUHAMMAD ASGHAR NAZAL[4]
[1,2,3)] Program Studi Statistika Fakultas MIPA Universitas Cenderawasih, Jayapura, Indonesia
[4)] Program Studi Sistem Informasi Fakultas MIPA Universitas Cenderawasih, Jayapura, Indonesia
e-mail : tasyairiani6@gmail.com

ABSTRAK

Pemilihan fitur adalah proses penting dalam pengembangan model pembelajaran mesin untuk mengidentifikasi fitur-fitur yang paling informatif dan relevan dalam dataset. Dua metode yang umum digunakan dalam pemilihan adalah metode Backward method dan Stepwise method. Pada penelitian ini, diterapkan Teknik Data Mining feature selection, untuk membandingkan kedua metode Feature Selection yaitu Backward Method dan Stepwise Method berdasarkan nilai accuracy. Hasil yang diperoleh dari perbandingan nilai akurasi Feature Selection yaitu Backward Method dan Stepwise Method menggunakan dataset Students Performance, kedua model ini sebanding. Dikatakan sebanding karena jika dilihat berdasarkan nilai akurasinya, nilai akurasi Backward Method dan Stepwise Method sama yaitu 0.61 atau 61%.

Kata Kunci : *Perbandingan, Feature Selection, Backward Method, Stepwise Method*

1. PENDAHULUAN

Berkembangnya teknologi informasi dalam bidang pengolahan data banyak membawa pengaruh positif bagi dunia pendidikan. Data mining dalam bidang pendidikan berfokus pada pengembangan eksplorasi tipe data yang unik, yang salah satu pemanfaatannya dapat dikembangkan dengan menggunakan metode dari algoritma data mining, *machine- learning* dan statistika. Salah satu metode dalam bidang data mining yang sering digunakan adalah metode klasifikasi.

Metode klasifikasi merupakan suatu metode yang digunakan untuk pengelompokan data kedalam kelas yang telah ditentukan. Teknik data mining yang digunakan pada penelitian ini adalah Feature Selection. Feature selection merupakan teknik pemilihan fitur, feature selection dapat digunakan untuk menyeleksi banyaknya fitur yang tidak relevan pada dataset. Teknik ini dilakukan untuk mengetahui subset fitur yang paling signifikan dari data set *Students Performance*. Sedangkan pengertian lain, feature selection atau seleksi fitur adalah suatu metode yang dapat digunakan untuk melakukan pengurangan dimensi terhadap dataset yang diberikan. Seleksi fitur dapat menyebabkan model yang dibangun lebih sederhana dan komprehensif, meningkatkan performa model, dan membantu untuk dapat memahami data. Feature Selection sering digunakan untuk pengurangan dimensi model. Feature Selection membantu mengurangi fitur domain, menghilangkan fitur yang berlebihan. Dengan cara ini akan membantu mempercepat proses pembelajaran/pemodelan. Pada penelitian ini digunakan dua metode feature selection, yaitu Backward Method dan Stepwise Method.

Backward method adalah salah satu metode feature selection yang bekerja dengan cara menghapus satu per satu fitur yang dianggap kurang relevan dalam model. Sementara itu, Stepwise method merupakan metode yang mengkombinasikan Forward Method dan Backward Method dengan menambahkan atau menghapus fitur satu per satu. Tujuan penelitian ini untuk membandingkan metode backward method dan stepwise method pada dataset students performance.

2. METODE PENELITIAN

Secara sistematis penelitian ini dibagi ke dalam beberapa tahapan penelitian yaitu pengumpulan data, perbandingan dan evaluasi.

2.1 Sumber Data

Data yang digunakan yaitu dataset Student's Performance yang diambil dari website kaggle. Dataset ini berisi informasi tentang kinerja siswa sekolah menengah atas dalam matematika, termasuk nilai dan informasi demografis mereka.

2.2 Pengumpulan Data

Berisi informasi tentang kinerja siswa sekolah menengah atas dalam matematika, termasuk nilai dan informasi demografis mereka. Data dikumpulkan dari tiga sekolah menengah atas di Amerika Serikat.

Jenis Kelamin	Pria / Wanita
Ras / etnis	Asia, afrika-amerika, hispanik, dll
Tingkat Pendidikan Orang Tua	Pendidikan terakhir orang tua atau wali siswa
Makan Siang	Ya / Tidak
Kursus Persiapan Ujian	Ya / Tidak
Nilai Matematika	Tes matematika standar
Nilai Membaca	Tes membaca standar
Nilai Menulis	Tes menulis standar

Kumpulan data ini dapat digunakan untuk berbagai pertanyaan penelitian terkait pendidikan orang tua atau kursus persiapan ujian terhadap kinerja siswa. Hal ini juga dapat digunakan untuk mengembangkan model pembelajaran mesin untuk memprediksi kinerja siswa berdasarkan demografi dan faktor lainnya.

2.3 Pre Processing Data

Berikut adalah tahapan dalam Pre Processing Data:

1. Pengecekan data duplikasi. Duplikasi dapat memengaruhi hasil analisis dan model, menyebabkan bias yang tidak diinginkan. Oleh karena itu, pembersihan duplikasi penting untuk memastikan keakuratan dan keandalan data. Pada dataset yang digunakan, data tersebut tidak memiliki data yang duplikasi.
2. Pengecekan Missing Value. Diawali dengan menghitung jumlah nilai yang hilang (missing value) dalam setiap kolom (fitur) dari DataFrame. Ini akan memberikan informasi tentang seberapa banyak nilai yang hilang dalam setiap kolom. Dalam dataset ini terdapat nilai yang hilang pada data lunch sebanyak 25. Setelah diketahui terdapat missing value, penulis menggunakan Imputation Missing Value untuk pengisian nilai yang hilang, untuk meningkatkan kualitas dan kegunaan dataset sehingga analisis atau pemodelan data dapat dilakukan dengan lebih baik.
3. Melakukan one-hot encoding. Mengubah variabel kategori menjadi bentuk yang dapat dimengerti oleh model machine learning. Model machine learning biasanya memerlukan input numerik, dan one-hot encoding adalah cara untuk mengatasi ini untuk variabel kategori. Ini mencegah model dari interpretasi keliru bahwa ada urutan atau peringkat antar kategori.

3. HASIL DAN PEMBAHASAN

Berdasarkan tujuan dari penelitian ini untuk mengetahui perbandingan menggunakan Backward Method dan Stepwise Method pada dataset Student's Performance. Dari hasil implementasi kedua metode, baik Backward Method (SBS) maupun Stepwise Method (SFFS), telah memilih subset variabel yang dianggap paling signifikan. Namun, perbandingan antara kedua metode tersebut mengungkapkan beberapa perbedaan penting:

1. Jumlah Variabel yang Dipilih :

   Dalam contoh backward, metode tersebut memilih 4 variabel terbaik yaitu: gender, parental level of education, lunch dan writing score. Sedangkan dalam contoh stepwise, metode tersebut memilih 3

variabel terbaik yaitu: gender, lunch dan writing score. Ini menunjukkan bahwa metode backward mungkin cenderung mempertahankan lebih banyak variable daripada metode stepwise.

2. Proses Pemilihan
   Metode backward secara interaktif menghapus satu fitur pada setiap langkahnya, sementara metode stepwise melakukan pencarian maju dan mundur secara bersamaan, menambahkan atau menghapus variabel berdasarkan kriteria yang ditentukan. Hal ini menunjukkan bahwa metode Stepwise dapat mempertimbangkan lebih banyak kombinasi variabel daripada metode backward.
3. Hasil Evaluasi Model
   Kedua metode menghasilkan model dengan akurasi yang sama, yaitu 0.61, menunjukkan bahwa kedua metode memberikan performa yang setara untuk dataset dan model yang digunakan.
4. Pemilihan Model
   Kedua metode menggunakan model klasifikasi Logistik Regresi yang sama dalam pembentukan model akhir, yang dapat mempengaruhi hasil evaluasi.

Dari perbandingan di atas, meskipun kedua metode memiliki perbedaan dalam subset variabel yang dipilih dan proses seleksinya, keduanya memberikan hasil akhir yang serupa dalam hal akurasi model. Pemilihan antara metode backward atau stepwise dapat bergantung pada preferensi spesifik, kompleksitas dataset, dan tujuan analisis yang diinginkan. Evaluasi tambahan dan eksperimen mungkin diperlukan untuk memahami perbedaan lebih lanjut antara kedua metode ini.

4. KESIMPULAN

Dalam penelitian ini, Backward Method dan Stepwise Method menghasilkan akurasi yang sama, yaitu 0,61 atau 61%. Hal ini menunjukkan bahwa keduanya mampu mempertahankan tingkat akurasi yang setara. Karena akurasi antara Backward method dan Stepwise method seimbang, pemilihan antara Backward Method dan Stepwise Method keduanya memberikan solusi parsimonius dalam statistik dengan memilih subset variabel yang menjelaskan variasi dataset tanpa menambahkan variabel yang tidak signifikan. Prinsip parsimoni dalam feature selection meminimalkan variabel tanpa mengorbankan kualitas prediksi, meningkatkan efisiensi komputasional, dan memudahkan interpretasi hasil, penting terutama bagi pemangku kepentingan non-statistik.

DAFTAR PUSTAKA

dkk Sarthika, “Analisis Profil Mahasiswa Politeknik Negeri Batam dengan Teknik Data Mining Asosiasi dan Clustering.” Vol. 8, no. 1, pp. 16–21 , 2016

B. Novianti, T. Rismawan, and S. Bahri, “Implementasi Data Mining dengan Algoritma C4. 5 untuk Penjurusan Siswa (Studi Kasus : SMA Negeri 1 Pontianak),” vol. 04, no. 3, 2016.

Chandani, Vinita., Romi Satria W., Purwanto., 2015. Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada analisis Sentiment Review Film. Journal of Intelligent Systems, Vol. 1, No.1, February 2015.

Guyon Isabelle dan A. Elisseeff, “An introduction to variable and Feature Selection,” Journal of Machine learning Research, Vol. 3, Edisi 7-8, 1157-1182, .

Husein Umar, 1998, Metodologi Penelitian : Aplikasi Dalam Pemasaran, PT. Gramedia Pustaka Utama, Jakarta.

Liu H, Motoda H, Setiono R. & Zhao Z. (2010). Feature Selection : An Eve Evolving Frontier in Data Mining", JMLR: Workshop and Conference Proceedings Vol.4, Publisher:Citeseer, pages 4-13.

Microsoft “MSDN : Feature Selection in Data Mining: Feature Selection in Analysis Services Data Mining”S. García, J. Luengo, and F. Herrera, “Feature Selection,” Intell. Syst. Ref. Libr., vol.72, no. 6, pp. 163-193, 2015, doi : 1 0.1007/978-3-319-10247-4_7.