e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

Implementasi Seleksi Fitur Pada Algoritma Klasifikasi Machine Learning Untuk Prediksi Penghasilan Pada Adult Income Dataset

Serafim Clara[1], Dhea Laksmi Prianto[2], Rizal Al Habsi[3], Ester Friscila Lumbantobing[4], Nurul Chamidah, S. Kom, M. Kom[5]
[1,2,3,4]Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok
serafim@upnvj.ac.id, dhealaksmi@gmail.com, rizalhabsi@upnvj.ac.id, esterfriscilalumbantobing@gmail.com, nurul.chamidah@upnvj.ac.id

Abstrak— Setiap orang mempunyai penghasilan yang berbeda-beda. Prediksi pada *Income Adult Dataset* menggunakan dua algoritma klasifikasi yaitu *Naive Bayes* dan *Random forest* untuk mengetahui nilai akurasi dengan mengimplementasikan seleksi fitur agar dapat mengetahui fitur yang paling berpengaruh. Eksperimen awal ialah melakukan analisis data. Target atribut yang digunakan adalah *Income*. Data memiliki 7% *missing value* pada beberapa atribut, oleh karena itu dibutuhkan beberapa proses *pre-processing* data sebelum melakukan tahap klasifikasi. Setelah melalui *pre-processing*, dilanjutkan dengan menerapkan seleksi fitur dengan memilih kandidat atribut yang terbaik dari hasil pertama dan selanjutnya dapat mengoptimalkan algoritma ini untuk memodelkan data terbaik. Hasil eksperimen menunjukkan bahwa *Naïve Bayes* adalah algoritma terbaik dari algoritma klasifikasi lainnya dengan hasil akurasi sebesar 84,51%.

Kata Kunci— seleksi fitur, *Naïve Bayes*, *Random forest*.

1 Pendahuluan

Klasifikasi merupakan salah satu teknik analisis pada *data mining*. Klasifikasi memiliki tujuan untuk memprediksi nilai keluaran dari fungsi dengan masukan baru setelah melalui proses *training*. Beberapa algoritma dengan teknik klasifikasi tersebut memiliki tingkat akurasi yang berbeda-beda. Akurasi ini digunakan sebagai pengatur dengan membandingkan nilai absolut. Semakin mendekati ukuran maka semakin tinggi nilai akurasinya. Atribut yang digunakan juga mempengaruhi nilai akurasi dan kompleksitas waktu dari masing-masing algoritma. Bukan hanya atribut, *record* data juga mampu mempengaruhi performa dari masing-masing algoritma

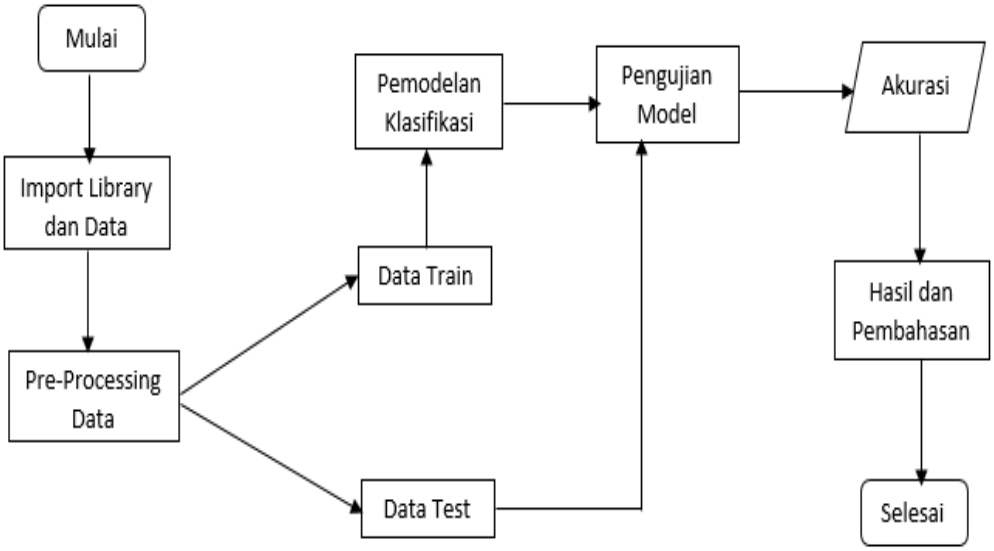
Data Adult Income menjadi salah satu contoh yang dapat diproses untuk melakukan klasifikasi. Data ini digunakan untuk mengklasifikasi seseorang yang memiliki pendapatan lebih dari 50.000 dolar per tahunnya dengan berdasarkan berbagai variabel yang diamati. Data ini berpengaruh untuk menjelaskan tingkat kemakmuran yang ada di masyarakat Amerika Serikat. Dalam data tersebut juga dijelaskan variabel yang diprediksi dapat mempengaruhi pendapatan sehingga dapat membantu pemerintah untuk memnentukan kebijakan bagi seluruh masyarajat Amerika Serikat.

Seleksi fitur sebagai sebuah teknik pengurangan dimensi menargetkan untuk memilih sebuah subset kecil dari sebuah fitur yang relevan dari fitur asli dengan menghilangkan fitur yang tidak relevan, redudansi fitur dan fitur yang memiliki *noisy* [1]. Penelitian ini mengukur kinerja dan membandingkan hasil pengukuran tingkat akurasi algoritma klasifikasi yaitu *Naive Bayes* dan *Random forest*. *Naïve Bayes classifier* merupakan klasifikasi probabilitas yang sederhana berdasarkan pada implementasi teorema Bayes dengan asumsi yang kuat (*naïve*) dan bebas (*independence*). Algoritma Naïve Bayes adalah sebuah pengelompokan kemungkinan sederhana yang mengkalkulasikan sebuah set probabilitas dengan menghitung frekuensi dan kombinasi nilai yang diberikan oleh dataset [2]. Random Forest adalah salah satu metode berbasis klasifikasi dan regresi dimana terdapat proses agregasi pohon keputusan [3]. *Random forest* tidak berkecendrungan untuk *overfitting* dan dapat diproses dengan cepat sehingga sangat memungkinkan untuk memproses *tree* yang diinginkan oleh pengguna. Dua algoritma

yang disebutkan tadi dapat menghasilkan tingkat keakuratan dan *f1-score* yang berbeda-beda dengan dataset yang sama. Berdasarkan latar belakang tersebut, maka penelitian kami akan membandingkan pengaruh seleksi fitur pada algoritma *Naive Bayes* dan *Random forest.*

2 Metode Penelitian

Proses pada penelitian ini untuk memprediksi mengenai *income* serta proses dalam mencari akurasi dari model-model yang digunakan. Dibawah ini adalah gambar dari tahapan yang akan digunakan.



Gambar 1. Metode Penelitian

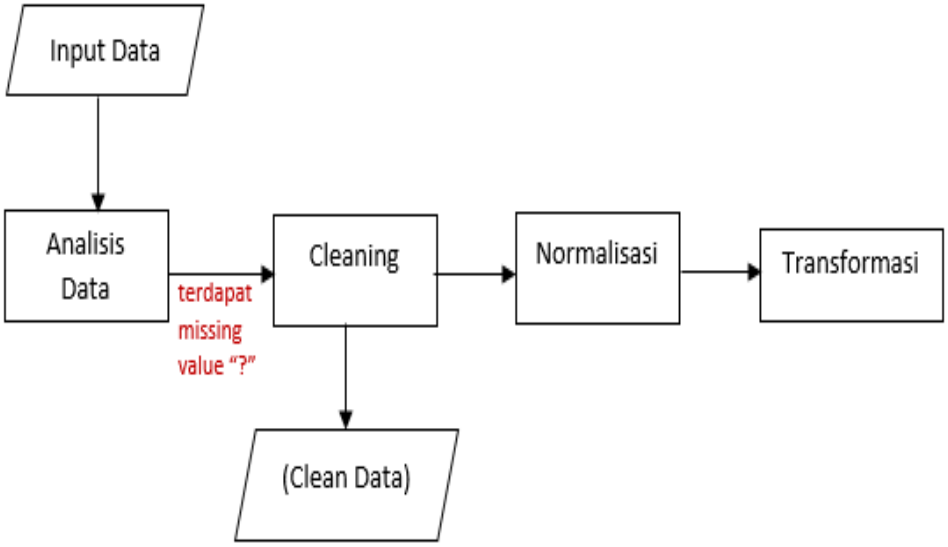
Pada gambar 1 terdapat metode penelitian dimulai dari import library dan data lalu melakukan tahapan *pre-processing* yang meliputi, data yang telah melewati tahapan *pre-processing* akan dibagi menjadi dua bagian mencakup *data train* dan *data test*, kemudian dmelakukan pemodelan klasifikasi menggunakan *data train* yang selanjutnya ialah pengujian model untuk data yang telah melewati tahap pemodelan klasifikasi maupun *data test* untuk melihat kedua akurasi dari data tersebut.

2.1 Data

Dataset yang digunakan merupakan dataset *adult income* yang berasal dari UCI, terdiri dari 48842 data dengan 14 fitur. Dataset ini berisi mengenai penghasilan yang lebih atau kurang dari $50.000 dalam waktu setahun dengan *missing value* sebesar 7%. Data yang digunakan dalam penelitian ini memiliki dimensi yang beragam sehingga dilakukan seleksi fitur agar dapat meningkatkan hasil akurasi.

Atribut yang ada pada dataset *Adult Income* yaitu Age, Workclass, Fnlwgt, *Education*, *education*-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income.

2.2 Pre-processing data



Gambar 2. *Pre-processing* data

Pre-processing data meliputi *cleaning*, yaitu menganti data atau menghilangkan data *noise* ataupun *missing value*, proses normalisasi untuk memodifikasi nilai dalam variabel sehingga kita dapat mengukurnya dalam skala umum atau rentang tertentu. Pada penelitian ini menggunakan *MinMax Normalization* lalu proses transformasi, mengubah data asli ke bentuk data tujuan agar mudah di proses.

Beberapa label dalam *income adult dataset* berbentuk kategorik diubah menjadi bentuk numerik.

Berikut data kategori yang diubah menjadi numerik:

- *Workclass*
- *Education*
- *marital-status*
- *occupation*
- *relationship*
- *race*
- *sex*
- *native-country*
- *income*

2.3 Split Data

Penentuan split data atau pembagian data yang dilakukan dengan rasio 70% data *training* dan 30% data *testing*.

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

2.4 Klasifikasi

Klasifikasi digunakan untuk memprediksi label kelas yang bersifat kategorikal dari sebuah atribut data yang diberikan. Algoritma klasifikasi mencari sebuah fungsi yang memberikan sebuah item kepada salah satu target yang telah ditentukan sebelumnya[6]. Teknik dari klasifikasi yang digunakan untuk dataset *Adult Income* meliputi:

2.4.1 *Naïve Bayes*

Klasifikasi dengan metode ini ialah mengambil asumsi atas penyederhanaan nilai atribut secara kondisional saling bebas jika diberikan nilai *output*. Metode ini digunakan untuk memprediksi probabilitas keanggotaan dari suatu kelas. Model ini merupakan model yang sederhana dan memiliki efisiensi yang cukup baik [7].

2.4.2 *Random Forest*

Random forest bekerja dengan beberapa *tree* atau pohon keputusan yang dimana masing-masing *tree* bergantung pada nilai piksel pada tiap *vector* yang diambil secara acak dan *independent* [5].

2.5 Evaluasi

Data *training* yang telah dibangun akan dilakukan tahap pengujian yang meliputi nilai akurasi dan nilai *F1 Score*. Nilai akurasi didapatkan dari prediksi benar untuk data positif dan negatif dari keseluruhan data. *F1 Score* adalah nilai yang menandakan jika model yang dibangun memiliki nilai presisi dan *recall* yang baik. Untuk memperoleh nilai akurasi, nilai *precision*, nilai *recall*, serta *f1 score* dapat menggunakan *Confusion Matrix* [8].

		Nilai Aktual	
		Positive	Negative
Nilai Prediksi	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

2.5.1. *Accuracy*

Accuracy adalah rasio yang memiliki prediksi nilai benar (nilai positif dan nilai negatif) berdasarkan keseluruhan data. Akurasi dapat menggambarkan keakuratan model klasifikasi yang digunakan. Nilai akurasi dapat diperoleh menggunakan persamaan berikut ini

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

2.5.2. *Precision*

Precision adalah rasio yang memiliki prediksi nilai benar positif jika dibandingkan dengan keseluruhan hasil yang diprediksi positif. *Precision* dapat menggambarkan keakuratan data yang diinginkan dengan hasil prediksi yang diperoleh model klasifikasi. Nilai *Precision* dapat diperoleh menggunakan persamaan berikut ini

$$precision = \frac{TP}{TP + FP} \quad (2)$$

2.5.3. *Recall*

Recall adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. *Recall* menggambarkan hasil dari model klasifikasi yang digunakan dalam menemukan kembali sebuah informasi. Nilai *Recall* dapat diperoleh menggunakan persamaan berikut ini

$$recall = \frac{TP}{TP+FN} \quad (3)$$

2.5.4. *F1 Score*

F1 Score adalah perhitungan kombinasi dari nilai *precision* dan nilai *recall* yang kemudian hasilnya disebut sebagai nilai pengukuran. *F1 Score* dapat diperoleh menggunakan persamaan berikut ini

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3 Hasil dan Pembahasan

Hasil eksperimen dengan menerapkan seleksi fitur pada dataset dengan menggunakan algoritma klasifikasi untuk mengetahui efektifitas reduksi data dan membandingkan dua algoritma klasifikasi dari akurasi yang diperoleh.

Metode seleksi fitur dilakukan dengan menggunakan *univariate statistics* untuk evaluasi apakah ada hubungan statistik yang signifikan dari setiap masukan fitur ke target (target variabel/*dependant variable*) yang akan kita prediksi. Fitur yang memiliki nilai kepercayaan yang tinggi yang akan digunakan untuk pemodelan. Perhitungan dilakukan dengan mengkalkulasi sebuah koefisien korelasi 2 seri dengan nilai p lalu menerapkan metode filter Karl Pearson dengan rumus:

$$r(x, y) = \frac{cov(x,y)}{\sigma x \sigma y} \quad (5)$$

Hasil nilai koefisien untuk korelasi metode seleksi fitur Karl Pearson adalah:

Tabel 1. Nilai Korelasi Pearson

Parameter	Nilai Absolut Pearson
Education	0.81196
Race	0.70844
Education Number	0.3328
Relationship	0.253402
Age	0.23704
Hours per Week	0.227199
Capital Gain	0.221034

Marital Status	0.192711
Capital Loss	0.148687
Occupation	0.049787
Native Country	0.020103
Workclass	0.015659
Fnlwgt	0.007264

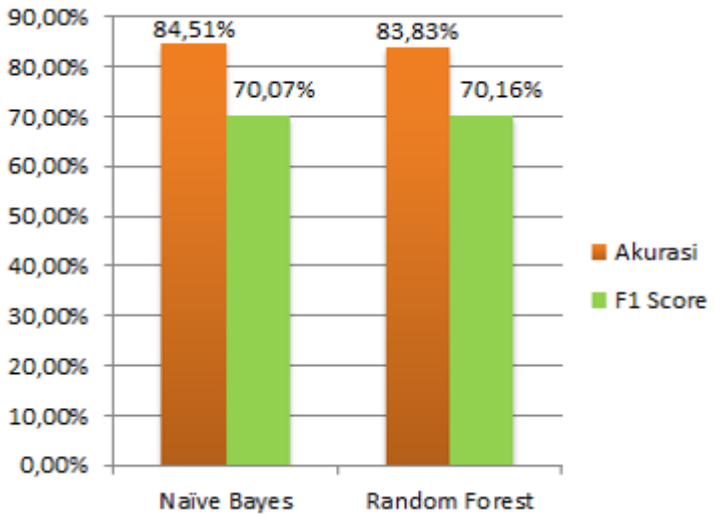
Yang selanjutnya akan dilakukan pemodelan berdasarkan 4 fitur terbaik, table 2 juga menunjukan bahwa fitur *Education Number of Years* (jumlah waktu dalam pendidikan) merupakan fitur yang paling relevan untuk menghasilkan nilai akurasi. Setelah dilakukan seleksi fitur, penelitian menggunakan empat fitur dengan persentase *Pearson's Correlation* yang paling tinggi. Dengan akurasi keseluruhan data fitur yang telah terseleksi sebesar 24.78%.

Tabel 2. Akurasi dengan seleksi fitur

Algoritma Klasifikasi	Akurasi	F1 Score
Naïve Bayes	84.51%	70.07%
Random forest	83.83%	70.16%

Tabel 2 merupakan tabel komparasi setiap algoritma yang telah menggunakan 4 fitur terbaik dalam proses seleksi fitur yang menunjukan bahwa algoritma *Naïve Bayes* memiliki nilai akurasi yang paling baik atau yang paling efektif jika dibandingkan dengan algoritma *Random forest.*

Nilai *F1 Score* atau rata-rata dari ketepatan dan rasio prediksi benar positif dibanding dengan keseluruhan data yang benar positif (*Recall*) Random Forest dengan 84.39%.



Gambar 3. Perbandingan sebelum seleksi fitur yang menunjukan hasil perbandingan dua algoritma dengan menerapkan seleksi fitur.

Mereduksi dimensi dengan *Feature Selection* dapat meningkatkan performa *predicted* model dan juga mengurangi waktu imputasi karena fitur yang tidak begitu relevan terhadap hasil akurasi dari klasifiksi berkurang. Dengan adanya perbandingan antara dua algoritma klasifikasi diatas dapat dilihat meskipun *Naïve Bayes* menghasilkan nilai keakuratan yang terbaik dibandingkan algoritma yang lain tetapi *Naïve Bayes* juga menghasilkan nilai *F1-Score* paling rendah.

4 Kesimpulan dan Saran

Kesimpulan yang dapat ditarik dari penelitian ini adalah bahwa dengan implementasi penerapan teknik seleksi fitur serta pemilihan algoritma yang berbeda untuk klasifikasi dengan menggunakan *dataset adult income* yang berasal dari UCI dapat mempengaruhi tingkat akurasi pada tahap evaluasi. Hasil yang didapatkan dari penelitian ini menunjukkan bahwa algoritma klasifikasi *Naïve Bayes* mendapatkan tingkat akurasi tertinggi dengan nilai 84.51% jika dibandingkan dengan algoritma *Random forest* yaitu dengan nilai akurasi 83.83%.

Namun sebaliknya, walaupun algoritma *Naïve Bayes* memiliki tingkat keakuratam yang tertinggi dibandingkan algoritma yang lain tetapi *Naïve Bayes* ini menghasilkan nilai *F1-Score* yang paling rendah. Hasil yang didapatkan dari penelitian ini menjelaskan bahwa nilai *F1-Score* yang tertinggi diraih *Random forest* dengan nilai 70.16% dan disusul dengan *Naïve Bayes* dengan nilai terendah yaitu 70.07%.

Adapun saran yang dapat peneliti berikan kepada pembaca untuk kedepannya ialah dengan mencoba menggunakan dua atau lebih seleksi fitur pada penelitian ini. Tujuannya ialah untuk membandingkan apakah dengan menggunakan seleksi fitur yang berbeda akan dapat mempengaruhi tingkat akurasi algoritma klasifikasi yang digunakan (*Naïve Bayes* dan *Random forest*).

Referensi

[1] Miao, J., & Niu, L. (2016). *A Survey on Feature Selection. Procedia Computer Science, 91*.

[2] Saritas, M. M. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2).

[3] Dhawangkhara, M., & Riksakomara, E. (2017). Prediksi Intensitas Hujan Kota Surabaya dengan Matlab menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya). Jurnal Teknik ITS, 6(1).

[4] Suyanto. (2018). Machine learning tingkat dasar dan lanjut, bandung, informatika

[5] Verma, A. (2018). Study and Evaluation of Classification Algorithms in Data Mining. *International Research Journal of Engineering and Technology*, *5*(8).

[6] Handian, D. (2017). gradDescentR 2.0: IMPLEMENTASI METODE GRADIENT DESCENT DAN VARIASINYA DALAM R PACKAGE (Doctoral dissertation, Universitas Pendidikan Indonesia).

[7] Tyas, R. A., Anggraini, M., Sulasiyah, I. A., & Aini, Q. (2020). Implementasi Algoritma Naïve Bayes Dalam Penentuan Rating Buku. *SISTEMASI: Jurnal Sistem Informasi*, *9*(3), 557-566.

[8] Rahman, M. F., Darmawidjadja, M. I., & Alamsah, D. (2017). Klasifikasi untuk diagnosa diabetes menggunakan metode bayesian regularization neural network (rbnn). *J. Inform*, *11*(1), 36.