JuTI "Jurnal Teknologi Informasi"
Fakultas Teknologi Informasi
Universitas Teknologi Digital Indonesia (d.h STMIK AKAKOM)
DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

Pengaruh Feature Selection Dan Feature Extraction Dalam Peningkatan Akurasi Klasifikasi Kebakaran Hutan

Andi Muhammad Rafli Armaya[1]

[1] Fakultas Teknik dan Ilmu Komputer, Universitas Indraprasta PGRI
Jl. Raya Tengah No. 80, Gedong, Pasar Rebo, Indonesia
[1]andirafli39@gmail.com (Corresponding author)

Disubmit: 22-08-23; diterima: 08-07-24; dipublikasikan: 05-08-24

Cara mengutip:
A.M.R. Armaya, 2024, "Pengaruh Feature Selection Dan Feature Extraction Dalam Peningkatan Akurasi Klasifikasi Kebakaran Hutan",*JuTI "Jurnal Teknologi Informasi"*, Vol. 3,No. 1, pp.13 – 23, DOI: 10.26798/juti.v3i1.1039

Ringkasan

Kebakaran hutan menjadi salah satu bencana yang menyebabkan kerugiaan yang sangat besar khususnya untuk lingkungan. Kebakaran hutan melepaskan sejumlah besar karbon dioksida, nitrogen oksida, belerang dioksida, dan gas rumah kaca lain yang mendorong terjadinya pemanasan global. Maka untuk menlakukan pencegahan dan meminimalisir dampak dari kebakaran hutan kita dapat melakukan data mining dengan cara klasifikasi. Dalam penelitian ini dilakukan beberapa algoritma antara lain decision tree (C4.5), logistic regression, neural network dan naïve bayes untuk meningkatan tingkat akurasi dan setelah mendapatkan algoritma yang baik, maka dilakukan feature selection dan feature extraction untuk meningkatkan akurasi algoritma tersebut..

Kata kunci: *Data Mining, Feature Selection, Feature Extraction, Klasifikasi, RapidMiner*

Abstract

Forest fires have become one of the disasters causing significant losses, especially to the environment. Forest fires release a substantial amount of carbon dioxide, nitrogen oxides, sulfur dioxide, and other greenhouse gases that contribute to global warming. To prevent and minimize the impact of forest fires, data mining can be utilized through classification methods. This study employs various algorithms such as decision tree (C4.5), logistic regression, neural network, and naive Bayes to enhance accuracy. After identifying effective algorithms, feature selection and feature extraction are performed to further improve their accuracy.

KeyWords: *Data Mining, Feature Selection, Feature Extraction, Klasifikasi, RapidMiner*

1. Pendahuluan

Kebakaran hutan merupakan masalah lingkungan yang serius dan dapat menyebabkan kerusakan ekosistem yang besar, termasuk hilangnya habitat satwa liar dan berkurangnya ketersediaan air bersih. Kerugian dan dampak negatif yang cukup besar akibat kebakaran hutan menyebabkan perlunya suatu usaha untuk pencegahan kebakaran hutan sejak dini. Melakukan sebuah prekdisi apakah kebakaran hutan akan terjadi atau tidak, tentu dapat membuat petugas melakukan pencegahan dini. Salah satu cara untuk memprediksi terjadinya kebakaran hutan adalah dengan menggunakan algoritma data mining.

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

Data mining merupakan proses yang memanfaatkan teknik matematika, statistik dan kecerdasan buatan untuk mengidentifikasi informasi atau pola – pola yang valid, baru, memiliki potensi bermanfaat dan bisa dipahami dari sekumpulan data yang besar[1].

Algoritma data mining memiliki banyak jenis, salah satunya adalah klasifikasi. Klasifikasi adalah proses pengelompokkan objek yang memiliki karakteristik atau ciri yang sama ke dalam beberapa kelas[2]. Terdapat banyak algoritma untuk melakukan klasifikasi, seperti naïve bayes, decision tree, k-Nearest Neighbors (k-NN), dan algoritma lainnya. Algoritma klasifikasi merupakan metode yang efektif untuk memprediksi potensi kebakaran hutan. Namun, akurasi prediksi dapat ditingkatkan dengan menggunakan teknik feature selection dan feature extraction yang tepat. Untuk mendapatkan algoritma dengan akurasi terbaik, maka beberapa algoritma klasifikasi akan dibandingkan dalam penelitian kali ini. Kemudian setelah mendapatkan algoritma dengan akurasi terbaik, selanjutnya akan diterapkan feature selection atau feature extraction pada algoritma tersebut untuk meningkatkan akurasinya. Meningkatkan akurasi prediksi potensi kebakaran hutan dapat membantu dalam mengambil tindakan preventif dan meminimalkan kerusakan yang disebabkan oleh kebakaran hutan.

Akurasi sendiri didefinisikan sebagai gerak atau kedekatan antara nilai yang terbaca dari alat ukur dengan nilai yang sebenarnya[3].

Ada banyak faktor yang dapat mempengaruhi potensi kebakaran hutan, termasuk cuaca, topografi, jenis vegetasi, dan kegiatan manusia. Oleh karena itu, perlu menggunakan teknik pemilihan fitur yang tepat untuk memilih fitur yang paling relevan dalam memprediksi potensi kebakaran hutan. *Feature selection* digunakan untuk mengurangi atribut atribut yang akan diproses oleh pengklasifikasi, mengurangi waktu eksekusi, dan meningkatkan akurasi. *Feature extraction* digunakan untuk mengekstrasi atribut sehingga mendapatkan nilai unik, penting, dan tidak duplikasi. Dengan kedua teknik ini diharapkan faktor faktor yang diproses merupakan faktor yang memang paling relevan sehingga akurasi prediksi dapat meningkat.

Sebagai seorang data scientist, ada sebuah tugas yaitu mencoba berbagai macam algoritma dan metode metode untuk menemukan sebuah cara yang akurat dan cocok untuk sebuah data tertentu. Pada penelitian kali ini akan dibandingkan 4 algoritma klasifikasi yaitu decision tree (C4.5), logistic regression, neural network dan naïve bayes untuk mencari algoritma mana yang paling baik untuk memprediksi kebakaran hutan pada data yang telah disediakan.

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang dikembangkan oleh Ross qiunlan. Beberapa pengembangan yang dilakukan pada C4.5 adalah sebagai antara lain bisa mengatasi missing value, bisa mengatasi continue data, dan pruning[4].

Regresi logistik adalah algoritma pembelajaran mesin yang paling terkenal setelah regresi linier. Dalam banyak hal, regresi linier dan regresi logistik serupa. Namun, perbedaan terbesar terletak pada apa yang mereka gunakan. Algoritma regresi linier digunakan untuk memprediksi / memperkirakan nilai tetapi regresi logistik digunakan untuk tugas klasifikasi[5].

Neural network merupakan sistem adaptif yang dapat merubah strukturnya untuk memecahkan masalah berdasarkan informasi eksternal maupun internal yang mengalir melalui jaringan tersebut. Secara sederhana neural network adalah sebuah alat pemodelan data statistik non-linear[6].

Algoritma Naïve Bayes merupakan algoritma klasifikasi yang menggunakan konsep dari probabilitas atau peluang. Dengan metode Naive Bayes terlebih dahulu mencari Nilai Probabilitas dan likelihood maksimum dari setiap atribut untuk masing-masing kelas[7].

Selain dibandingkan keempat algoritmat tersebut juga diaplikasikan dengan 2 teknik feature selection, yaitu forward selection dan backward elimination, serta 2 teknik feature extraction, yaitu *Principal Component Analysis* (PCA) dan *Self-Organizing Map* (SOM).

Feature selection adalah salah satu teknik penting dan sering digunakan dalam pre-processing. Teknik ini mengurangi jumlah fitur yang terlibat dalam menentukan suatu nilai kelas target, mengurangi fitur yang tidak relevan, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target[1]. Forward selection dimulai dengan fitur himpunan kosong lalu menambahkan fitur yang terpakai pada putaran pertama, semua fitur dievaluasi masing-masing. Salah satu fitur ditambahkan pada fitur himpunan yang merupakan bagian dari fitur sebelumnya dan juga fitur yang baru dibuat, lalu dievaluasi kembali. Untuk mengurangi jumlah evaluasi, hanya subset fitur terbaik yang disimpan[8]. Sedangkan backward elimination merupakan suatu metode yang memiliki fungsi untuk mengoptimalkan kinerja suatu model dengan sistem kerja pemilihan mundur. Pemilihan variabel dilakukan dengan cara pemilihan kedepan yakni menguji semua variabel kemudian menghapus variabel-variabel yang dianggap tidak signifikan[9].

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

Kemudian feature extraction adalah fase penting dalam identifikasi karena setiap huruf mempunyai keunikan tersendiri sehingga membedakan dirinya dari huruf yang lain. Feature extraction bertujuan untuk mendapatkan karakteristik suatu karakter yang membedakannya dari karakter lain yang disebut feature[10]. PCA termasuk dalam bidang multivariate analysis pada ilmu statistik. multivariate analysis secara sederhana dijelaskan sebagai metode yang berhubungan dengan variable dalam jumlah besar pada satu atau banyak percobaan[11]. Sedangkan SOM merupakan suatu jaringan konohen yang tidak membutuhkan suatu pengawasan khusus, karenanya diberi nama self-organizing. Kata maps berarti bahwa metode ini menggunakan map dalam pembobotan input data. Tiap node dalam jaringan SOM berusaha untuk menjadi seperti input yang telah diberikan pada jaringan tersebut[12].

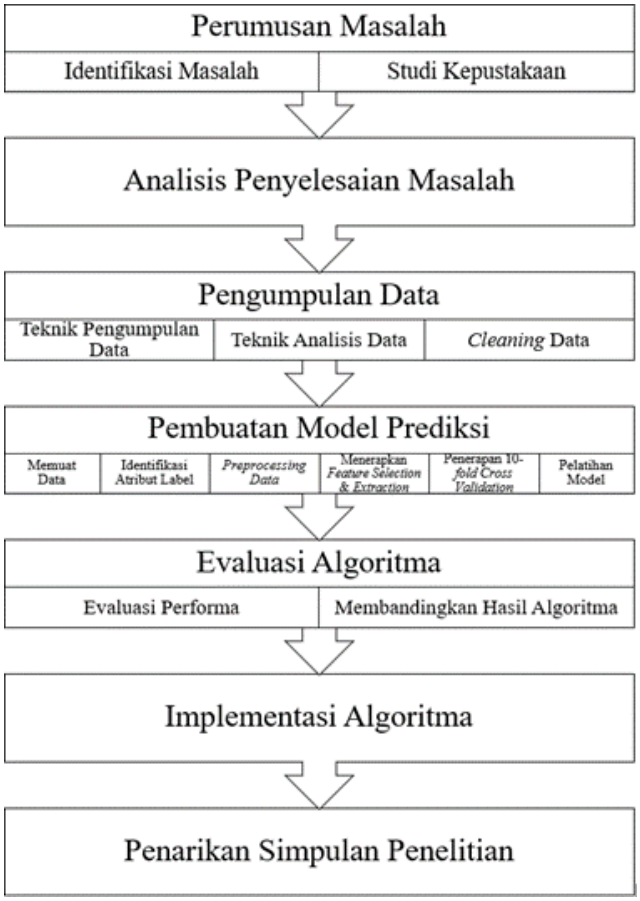
Untuk melakukan penelitian ini, alat yang digunakan adalah software RapidMiner. RapidMiner sendiri adalah alat analisis data yang dapat digunakan untuk membangun model prediksi potensi kebakaran hutan dengan menggunakan algoritma klasifikasi yang berbeda. Selain dapat membangun model dengan algoritma klasifikasi, RapidMiner juga dapat menerapkan feature selection dan feature extraction ke dalam model. Oleh karena itu, RapidMiner dapat digunakan untuk mengembangkan model prediksi yang lebih akurat dan efektif. Kemudian dataset yang digunakan adalah dataset yang berasal dari web open source, UCI Repository.

2. Metode Penelitian

Dalam waktu kurang lebih 4 bulan, yaitu dari bulan April sampai bulan Juli 2023. Dikarenakan penelitian menggunakan data open-source yang dapat diakses dengan komputer, maka Penulis melakukan penelitian di rumah penulis.

Tahapan penelitian terdiri dari 7 tahap yaitu perumusan masalah yang terdiri dari identifikasi masalah dan studi kepustakaan. Kemudian di lanjutkan dengan analisis penyelesaian masalah. Setelah selesai analisis penyelesaian masalah, tahap selanjutnya adalah pengumpulan data dan analisis data untuk menyelesaikan data, serta membersihkan data yang sudah didapatkan. Setelah selesai menganalisis data dan data sudah siap diolah, maka selanjutnya adalah pembuatan model menggunakan RapidMiner yang terdiri dari 6 langkah, yaitu memuat data ke RapidMiner, mengidentifikasi atribut yang menjadi label, melakukan preprocessing data, menerapkan feature selection dan extraction kepada data, penerapan 10-fold cross validation, serta pelatihan model. Dilanjutkan dengan evaluasi dan membandingkan hasil algoritma. Langkah terakhir adalah implementasi algoritma dan penarikan kesimpulan penelitian. Ketujuh tahap itu dijabarkan sebagai berikut:

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118



Gambar 1. Workflow Penelitian

1. Perumusan Masalah

   (a) Identifikasi Masalah
   Prediksi potensi kebakaran hutan yang akurat merupakan tantangan dalam upaya mitigasi dan pengendalian kebakaran, karena melibatkan data yang kompleks dan memerlukan metode yang tepat dan akurat

   (b) Studi Kepustakaan
   Studi kepustakaan dilakukan dengan mencari jurnal jurnal yang memiliki kesamaan contoh kasus, yaitu menggunakan feature selection dan feature extraction dalam penelitiannya. Beberapa jurnal tersebut sudah dicantumkan pada bab 2, sub-bab "penelitian yang relevan".

2. Analisis Penyelesaian Masalah
   Untuk mendapatkan sebuah metode yang tepat dan akurat seorang data scientist melekakuan trial and error. Melakukan beberapa kali percobaan dengan berbagai metode yang berbeda. Pada penelitian ini setidaknya terdapat 16 model prediksi yang dibentuk, terdiri dari 4 algoritma klasifikasi yaitu decision tree (C4.5), logistic regression, neural network dan naïve bayes. Kemudian keempat algoritma tersebut diaplikasikan dengan *feature selection* yang terdiri dari *Forward Selection* dan *Backward Elimination*, serta *Feature Extraction* yang terdiri dari *Principal Component Analysis* (PCA) dan Self Organizing Map (SOM).

3. Pengumpulan Data

   (a) Teknik Pengumpulan Data
   Teknik pengumpulan data yang digunakan dalam penelitian ini adalah studi literatur. Studi literatur merupakan teknik pengumpulan data dengan cara mencari, membaca, dan mempelajari buku buku, serta melakukan riset melalui internet sebagai literatur yang dapat mendukung dalam penyusunan dan penulisan skripsi. Studi literatur yang dilakukan dalam penelitian kali ini mencari dan mempelajari data di web open source, yaitu UCI repository (https://archive.ics.uci.edu/). Data

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

yang kemudian terpilih adalah "Algerian Forest Fires Dataset" (https://doi.org/10.24432/C5KW4N). Data tersebut berisi 2 sheet yaitu hutan area Sidi-Bel Abbes dan area Bejaia. Adapun spesifikasi data algetrian forest fires adalah terdiri dari 122 baris setiap tabel dengan 14 kolom. Kolom tersebut antara lain "Day" yang berisikan data tanggal, "Month" yang terdiri data bulan dari juni ke september, "Year" berisi data tahun, yaitu 2012. Kemudian ada kolom Temperature berisikan data suhu di angka 22 - 42 dalam satuan celsius, RH (Relative Humidity) adalah data Kelembapan udara di angka 21 - 90 dalam satuan persen, WS (Wind Speed) adalah data Kecepatan Angin di angka 6 - 29 dalam satuan Km/h, Rain berisikan data curah hujan di angka 0 - 16.8 dalam satuan mm. Kemudian ada system peringkat kebakaran hutan, yang terdiri dari "Fine Fuel Moisture Code" (FFMC) index dari sistem FWI di angka 28.6 - 92.5, "Duff Moisture Code" (DMC) index dari sistem FWI di angka 1.1 - 65.9, "Drought Code" (DC) index dari sistem FWI di angka 7 - 220.4, "Initial Spread Index" (ISI) dari sistem FWI di angka 0 - 18.5, "Buildup Index" (BUI) index dari sistem FWI di angka 1.1 - 68, "Fire Weather Index" (FWI) Index di angka 0 - 31.1. Terakhir ada kolom "Classess" yang berisika data hutan tersebut terjadi kebakaran atau tidak. Kolom "Classess" ini lah yang nanti akan menjadi label untuk prediksi.

(b) Teknik Analisis Data
Teknik yang dilakukan pada penelitian ini adalah klasifikasi. Pada tahap analisis penyelesaian masalah, berbagai algoritma klasifikasi dapat diimplementasikan dan dievaluasi untuk prediksi potensi kebakaran hutan. Beberapa algoritma yang umum digunakan meliputi decision tree (C4.5), logistic regression, neural network dan naïve bayes. Analisis ini melibatkan pembagian data menjadi data pelatihan dan data pengujian, serta pengukuran performa model klasifikasi menggunakan metrik evaluasi seperti akurasi. Keempat algoritma tersebut juga akan diaplikasikan dengan teknik feature selection dan feature extraction, yaitu Forward Selection dan Backward Elimination, Principal Component Analysis (PCA) dan Self Organizing Map (SOM). Selain menggunakan metode klasifikasi, penelitian ini juga menggunakan k-fold cross-validation, lebih tepatnya 10-fold cross-validation. Teknik ini digunakan untuk menguji kehandalan dan generalisasi model klasifikasi. Dengan menggunakan metode cross-validation seperti k-fold cross-validation, data dapat dibagi menjadi subset yang saling tumpang tindih untuk melatih dan menguji model secara iteratif. Hal ini membantu menghindari bias dan overfitting, serta memberikan perkiraan yang lebih dapat diandalkan tentang performa model. Kemudian yang terakhir adalah evaluasi performa. Selama analisis penyelesaian masalah, penting untuk mengukur performa model klasifikasi dengan menggunakan metrik evaluasi yang relevan. Metrik yang digunakan pada penelitian ini adalah akurasi.

(c) Cleaning Data
Data yang didapatkan terbagi menjadi 2 region, yaitu Sidi-Bel Abbes dan Bejaia serta memiliki kolom tanggal, bulan dan tahun yang terpisah. Memproses data yang berisi 2 tabel dalam satu sheet tidak bisa dilakukan dalam RapidMiner, oleh karena itu hal yang dilakukan sebelum memasukan data ke RapidMiner adalah menjadikan 2 tabel tersebut menjadi 1 sheet dengan membuat kolom baru dengan nama "Region". Daerah Bejaia akan diberi nomor 1 dan Sidi-Bel Abbes akan diberi nomor 2 dalam kolom tersebut. Kemudian atribut tanggal, bulan dan tahun akan dijadikan 1 dalam kolom "Date" yang sudah memuat ketiga kolom tersebut supaya data lebih mudah dibaca dan mudah diolah dengan RapidMiner.

4. Pembuatan Model Prediksi Berikut adalah langkah-langkah dalam membuat model prediksi dengan menggunakan rapidminer:

(a) Memuat Data
Data yang telah dikumpulkan untuk penelitian, yang terkait dengan potensi kebakaran hutan, akan dimuat ke dalam lingkungan RapidMiner. Data dapat berupa file CSV, Excel, atau format lainnya yang didukung oleh RapidMiner. Pada penelitian kali ini format datanya adalah CSV.

(b) Identifikasi Atribut Target
Atribut target atau label yang akan diprediksi dalam penelitian ini adalah potensi kebakaran hutan. Pada tahap ini, atribut target akan diidentifikasi dan dipisahkan dari atribut-atribut lainnya dalam dataset. Dalam data penelitian nama kolom yang akan dijadikan atribut adalah Classes, kolom tersebut berisi fire dan not fire.

(c) Preprocessing Data
Tahap preprocessing data melibatkan pengelolaan dan pemrosesan data untuk memastikan kualitasnya. Ini mungkin meliputi penghapusan missing values, penanganan outliers, normalisasi data, atau transformasi data jika diperlukan.

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

(d) *Feature Selection* dan *Feature Extraction*
Pada langkah ini akan diterapkan pemilihan fitur (*feature selection*) atau ekstraksi fitur (*feature extraction*) sebelum pembentukan model. Pemilihan fitur melibatkan pemilihan subset fitur yang paling relevan dan informatif untuk prediksi potensi kebakaran hutan. Sementara itu, ekstraksi fitur melibatkan penggabungan atau transformasi fitur yang ada menjadi representasi yang lebih baik. Dikarenakan *feature extraction* tidak dapat mengakstrak tipe data date, oleh karena itu pada tahap ini diputuskan bahwa atribut date dihapus dari tabel. Pada penelitian ini *feature selection* yang digunakan adalah *Forward Selection* dan *Backward Elimination*. Dalam melakukan Forward Selection hal pertama dilakukan adalah memisahkan atribut dan label. Tahapan melakukan *forward selection* adalah:

i. Pisahkan data menjadi atribut (fitur) dan label (kelas). Atribut adalah kolom "Region", "Temperature", "RH", "Ws", "Rain", "FFMC", "DMC", "DC", "ISI", "BUI", dan "FWI", sedangkan label adalah kolom "Classes"
ii. Jika menggunakan python maka buat sebuah variabel kosong yang nanti digunakan untuk menyimpan atribut yang tepilih. Dalam penelitian ini variabel tersebut diberi nama "selected_features".
iii. Mulai iterasi dengan 1 atribut yang ada dan di setiap iterasi tambahkan satu atribut baru untuk mencari atribut terbaik yang meningkatkan performa model.
iv. Mulai algoritma klasifikasi, seperti Logistic Regression pada 1 atribut yang dipilih dan lakukan 10-fold cross-validation untuk menilai performa model saat menambahkan setiap atribut baru kemudian evaluasi hasil akurasi.
v. Atribut yang mempengaruhi peningkatan performa akurasi secara signifikan dapat ditambahkan ke variabel kosong sebelumnya, yaitu "selected_features"
vi. Ulangi langkah 4 dan 5 sampai tidak ada atribut lagi yang memberikan peningkatan akurasi secara signifikan. Setelah selesai, maka "selected_features" akan berisi atribut yang memberikan peningkatan akurasi secara signifikan.

Sedangkan untuk melakukan backward elimination berkebalikan dengan forward section, yaitu membuat model dengan seluruh atribut lengkap terlebih dahulu, kemudian menghapus satu per satu atribut yang tidak mempengaruhi akurasi secara signifikan. Tahapannya adalah sebagai berikut:

i. Pisahkan data menjadi atribut (fitur) dan label (kelas). Atribut adalah kolom "Region", "Temperature", "RH", "Ws", "Rain", "FFMC", "DMC", "DC", "ISI", "BUI", dan "FWI", sedangkan label adalah kolom "Classes".
ii. Buat model awal dengan seluruh atribut yang ada.
iii. Mulai lakukan iterasi dengan seluruh atribut dan setiap iterasi hapus atribut yang tidak mempengaruhi peningkatan akurasi secara signifikan.
iv. Mulai algoritma klasifikasi, seperti Logistic Regression pada semua atribut dan lakukan *10-fold cross-validation* untuk menilai performa model saat menggunakan atribut-atribut saat ini.
v. Hapus satu atribut dengan performa terburuk dari model saat ini. Atribut yang dihapus adalah Atribut yang memiliki kontribusi paling rendah dalam performa model.
vi. Latih ulang model setelah menghapus satu atribut, kemudian evaluasi hasil akurasi model baru.
vii. Ulangi langkah 5 dan 6 hingga performa model tidak lagi meningkat secara signifikan atau telah mencapai ambang batas yang ditentukan.

Selain *feature selection* pada penelitian ini menerapkan *feature extraction* yaitu *Principal Component Analysis* (PCA) dan *Self Organizing Map* (SOM). Tahapan untuk melakukan *Component Analysis* (PCA):

i. Mulai dengan meniapkan data kebakaran hutan yang telah dicleaning sebelumnya. Pastikan data sudah diatur dalam bentuk tabel dengan baris mewakili sampel dan kolom mewakili atribut.
ii. Hitung rata-rata dari setiap fitur dan lakukan normalisasi data jika diperlukan.
iii. Hitung matriks kovariansi dari data.
iv. Lakukan PCA dengan menghitung eigenvalue dan eigenvector dari matriks kovariansi.
v. Pilih sejumlah komponen utama berdasarkan eigenvalue terbesar untuk mengurangi dimensi data.

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

vi. Proyeksikan data ke dalam sistem koordinat baru berdasarkan komponen utama yang dipilih.
vii. Buat scatter plot atau grafik lain untuk memvisualisasikan data dalam dimensi yang lebih.

Selanjutnya adalah tahapan untuk melakukan teknik *Self Organizing Map* (SOM):

i. Tentukan ukuran grid SOM (misalnya 2x2 atau 5x5) untuk menyusun neuron-neuron dalam grid.
ii. Inisialisasi bobot untuk setiap neuron dalam grid SOM secara acak.
iii. Lakukan iterasi untuk mengupdate bobot neuron berdasarkan jarak dari input data (unsupervised learning).
iv. 4) Proyeksikan data ke dalam grid SOM untuk menemukan neuron pemenang (neuron dengan bobot terdekat dengan data input).
v. Visualisasikan peta SOM dengan menggunakan warna atau marker berbeda untuk setiap kelompok data yang terbentuk.

(e) Penerapan *10-Fold Cross Validation*
K-Fold Cross Validation adalah teknik evaluasi model data mining untuk tujuan memperkirakan kenerja atau akurasi model secara lebih andal dengan memanfaatkan seluruh data yang tersedia untuk training dan testing. K yang biasa digunakan adalah 5 atau 10. K = 10 telah diakui secara umum dalam literatur dan praktik machine learning sebagai nilai yang baik untuk cross-validation. Hal ini membuatnya menjadi pilihan yang populer. Selain populer K = 10 juga mudah dipahami oleh para peneliti dan praktisi. Contoh penggunaannya adalah membagi data menjadi K yang dipilih. Misal ada 100 data dan K = 10, maka iterasi pertama 10 data teratas menjadi data testing dan 90 data selanjutnya data training, iterasi kedua data ke-11 sampai ke- 20 data testing dan 90 data sisanya menjadi training. Terus berlanjut sampai 10 kali iterasi karena K = 10 atau ke-100 data tersebut telah menjadi data training maupun data testing.

(f) Pelatihan Model
Setelah algoritma dikonfigurasi, model akan dilatih menggunakan data pelatihan. Ini melibatkan proses belajar algoritma dari data pelatihan dan penyesuaian model untuk melakukan prediksi potensi kebakaran hutan. Algoritma yang digunakan ada 4 buah, yaitu C4.5, logistic regression, neural network dan naïve bayes. Dalam melatih model pada penelitian ini digunakan software RapidMiner, oleh karena itu data perlu dimuat ke dalam rapidminer. Proses pemilihan label untuk prediksi, preprocessing data, penerapan *Forward Selection*, *Backward Elimination*, Principal *Component Analysis* (PCA), *Self-Organizing Map* (SOM) serta *10-fold cross validation*.

(g) Evaluasi dan Hasil Algoritma

i. Evaluasi Performa
Setelah pelatihan model selesai, performa model akan dievaluasi menggunakan metrik evaluasi yang relevan seperti akurasi. Evaluasi ini akan memberikan informasi tentang seberapa baik model dapat memprediksi potensi kebakaran hutan.
ii. Perbandingan Algoritma
Setelah evaluasi performa, hasil dari setiap algoritma klasifikasi akan dibandingkan untuk menentukan algoritma mana yang memberikan hasil yang paling baik dalam memprediksi potensi kebakaran hutan. Algoritma yang dibandingkan adalah C4.5, logistic regression, neural network dan naïve bayes.

(h) Implemenasi Algoritma
Algoritma yang menghasilkan model dengan akurasi terbaik saat dibuat dengan menggunakan RapidMiner akan dicoba untuk diimplementasikan dengan data dummy yang peniliti buat. Implementasi tersebut akan menggunakan bahasa program python dengan Google Collaboratory.

(i) Penarikan Simpulan Penelitian
Setelah seluruh langkah langkah di atas selesai dikerjakan dan mendapatkan algoritma dengan akurasi terbaik, maka langkah terakhir adalah membuat kesimpulan dari hasil penelitian.

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

3. Hasil dan Pembahasan

3.1. Pembentukan Model

Gambar 2. Model RapidMiner

Seperti pada Gambar 2, sebelum melakukan klasifikasi, langkah pertama adalah kolom "tanggal" pada data yang sudah rapih dan dimasukan ke dalam RapidMiner dihapus terlebih dahulu karena feature extraction tidak dapat mengekstrak tipe data "date".

Setelah itu maka data sudah dapat diaplikasikan dengan feature selection dan feature extraction. Feature selection yang digunakan adalah Forward Selection dan Backward Elimination. Kemudian feature extraction yang digunakan adalah *Principal Component Analysis* (PCA) dan *Self-Organizing Map* (SOM). Supaya data yang digunakan dalam pengujian konsisten, maka gunakan operator "multiply" agar setiap data yang masuk ke dalam model dalam keadaan sama.

Kemudian, setelah data sudah rapih dan telah diterapkan feature selection dan feature extraction, maka langkah selanjutnya membuat model dengan melakukan 4 buah percobaan dengan algoritma Decision Tree (C4.5), Logistic Regresion, Neural Network, Naïve Bayes.

3.1.1. Hasil

Hasil dari model pada gambar 2 dapat dilihat pada Tabel 1 dibawah ini:

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

Tabel 1. Hasil Akurasi Model

Algoritma	Akurasi
Decision Tree (C4.5)	
Forward Selection	97.98%
Backward Elimination	98.35%
Principal Component Analysis (PCA)	86.05%
Self-Organizing Map (SOM)	80.70%
Logistic Regresion	
Forward Selection	98.8%
Backward Elimination	98.37%
Principal Component Analysis (PCA)	84.88%
Self-Organizing Map (SOM)	57.72%
Neural Network	
Forward Selection	98.35%
Backward Elimination	98.38%
Principal Component Analysis (PCA)	87.32%
Self-Organizing Map (SOM)	84.75%
Naïve Bayes	
Forward Selection	97.98%
Backward Elimination	96.32%
Principal Component Analysis (PCA)	84.45%
Self-Organizing Map (SOM)	79.10%

Dapat dilihat bahwa yang memiliki akurasi tertinggi adalah *logistic regression* yang sudah diaplikasikan dengan *forward selection*. Oleh karena itu maka model ini akan diimplementasikan ke dalam python.

3.1.2. Implementasi Python

Implementasi dari model dalam Python dapat dilihat pada Gambar 3 berikut, dengan sebagian data yang digunakan dalam penelitian ini pada Tabel 2.

```
[ ]  X = data.drop(columns=['Classes'])  # Memisahkan atribut dari label
     y = data['Classes']  #Label
```

```
 #Mendefinisikan fungsi forward_selection_with_cv dengan parameter X, y, significance_level, dan cv.
def forward_selection_with_cv(X, y, significance_level=0.05, cv=10):
    features = list(X.columns)
    selected_features = []
    best_score = 0

    while len(features) > 0:
        remaining_features = list(set(features) - set(selected_features))
        scores = []

        for feature in remaining_features:
            X_temp = X[selected_features + [feature]]
            model = LogisticRegression()
            score = np.mean(cross_val_score(model, X_temp, y, cv=cv))
            scores.append((feature, score))

        # Memilih fitur dengan score terbaik
        best_feature, max_score = max(scores, key=lambda x: x[1])
        if max_score > best_score:
            best_score = max_score
            selected_features.append(best_feature)
        else:
            break #berhenti

    return selected_features #Mengembalikan nilai ke "selected_features"
```

```
[ ]  selected_features = forward_selection_with_cv(X, y)
     print("Fitur yang dipilih:")
     print(selected_features)
```

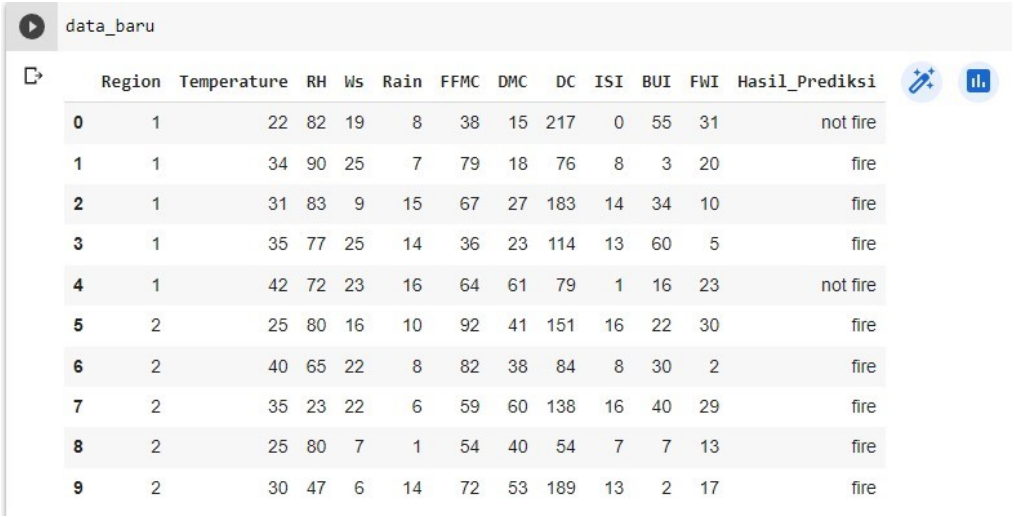
Gambar 3. Pembentukan Model di Python

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

Tabel 2. Data Pengujian Model

Region	Temperature	RH	Ws	Rain	FFMC	DM C	DC	IS I	BU I	FWI
1	22	82	19	8	38	15	217	0	55	31
1	34	90	25	7	79	18	76	8	3	20
1	31	83	9	15	67	27	183	14	34	10
1	35	77	25	14	36	23	114	13	60	5
1	42	72	23	16	64	61	79	1	16	23
2	25	80	16	10	92	41	151	16	22	30
2	40	65	22	8	82	38	84	8	30	2
2	35	23	22	6	59	60	138	16	40	29
2	25	80	7	1	54	40	54	7	7	13
2	30	47	6	14	72	53	189	13	2	17

Dengan sebagian hasil pengujian dapat dilihat pada Gambar 4.



data_baru

	Region	Temperature	RH	Ws	Rain	FFMC	DMC	DC	ISI	BUI	FWI	Hasil_Prediksi
0	1	22	82	19	8	38	15	217	0	55	31	not fire
1	1	34	90	25	7	79	18	76	8	3	20	fire
2	1	31	83	9	15	67	27	183	14	34	10	fire
3	1	35	77	25	14	36	23	114	13	60	5	fire
4	1	42	72	23	16	64	61	79	1	16	23	not fire
5	2	25	80	16	10	92	41	151	16	22	30	fire
6	2	40	65	22	8	82	38	84	8	30	2	fire
7	2	35	23	22	6	59	60	138	16	40	29	fire
8	2	25	80	7	1	54	40	54	7	7	13	fire
9	2	30	47	6	14	72	53	189	13	2	17	fire

Gambar 4. Hasil Pengujian Data

4. Simpulan

Dari penelitian ini, dapat disimpulkan bahwa menggunakan kombinasi metode feature selection dengan forward selection, algoritma logistic regression, dan evaluasi dengan 10-fold cross validation berhasil meningkatkan akurasi prediksi potensi kebakaran hutan. Melalui eksperimen dan analisis, kita dapat mengidentifikasi atribut-atribut yang paling berpengaruh dalam memprediksi potensi kebakaran hutan. Hasil akhir model Logistic Regression yang telah dilatih menunjukkan tingkat akurasi yang tinggi, yaitu 98,8% yang mengindikasikan model tersebut memiliki kemampuan yang sangat baik dalam mengklasifikasikan kebakaran hutan.

Pustaka

[1] A. Pangestu, R. Dedy, S. T. Wijaya, E. Hernawati, and M. Kom, "Aplikasi Pengolahan Data Prediksi Kemiskinan Berbasis E-Commerce Menggunakan Decision Tree Dan Wrapper Feature Selection Application Based of E-Commerce Poverty Prediction Data Processing Decision Tree and Wrapper Feature Selection," pp. 1729–1740, 2020.

[2] R Ikhsan and A Turmudin Zy, "Penentuan Jadwal Overtime Dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5," *Jurnal Sains Komputer dan Informatika (J-SAKTI)*, pp. 694–702, 2020.

[3] Milda, "AKURASI TAKARAN DALAM JUAL BELI BERAS di PASAR SALUDONGKA KECAMATAN PAKUE UTARA KABUPATEN KOLAKA UTARA," Ph.D. dissertation, 2017. [Online]. Available: http://repository.iainpalopo.ac.id/id/eprint/2206/

DOI: 10.26798/juti.v3i1.1039 vol.3, No.1, Agustus, e-ISSN: 2962-4118

[4] E. P. K. Orpa, E. F. Ripanti, and T. Tursina, "Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision Tree C4.5," *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, vol. 7, no. 4, p. 272, 2019.

[5] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 6, no. 3, p. 280, 2020.

[6] S. Dewi, "Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan," *Techno Nusa Mandiri*, vol. 13, no. 1, pp. 60–66, 2016.

[7] S. Karthika and N. Sairam, "A Naïve Bayesian classifier for educational qualification," *Indian Journal of Science and Technology*, vol. 8, no. 16, 2015.

[8] M. R. Fanani, "Algoritma Naïve Bayes Berbasis Forward Selection Untuk Prediksi Bimbingan Konseling Siswa," *Jurnal DISPROTEK*, vol. 11, no. 1, pp. 13–22, 2020.

[9] A. Bode, "K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika," *ILKOM Jurnal Ilmiah*, vol. 9, no. 2, pp. 188–195, 2017.

[10] J. Sitorus, "Perancangan Aplikasi Pengenalan Pola Huruf Aksara Batak Toba Menerapkan Metode Direction Feature Extraction (DFE)," *JURIKOM (Jurnal Riset Komputer)*, vol. 2, no. 6, pp. 48–55, 2015.

[11] A. Suryadi, "Sistem Pengenalan Wajah Menggunakan Metode Principal Component Analysis (PCA) Dengan Algoritma Fuzzy C-Means (FCM)," *Mosharafa: Jurnal Pendidikan Matematika*, vol. 4, no. 2, pp. 58–65, 2015.

[12] N. N. Halim and E. Widodo, "Clustering dampak gempa bumi di Indonesia menggunakan kohonen self organizing maps," *Prosiding SI MaNIS (Seminar Nasional Integrasi Matematika dan Nilai Islami)*, vol. 1, no. 1, pp. 188–194, 2017. [Online]. Available: http://conferences.uin-malang.ac.id/index.php/SIMANIS/article/view/62