Pendekatan Penambangan Data untuk Memprediksi Kebakaran Hutan Menggunakan Data Meteorologi

Paulo Cortez1 dan An´ýbal Morais1

Departemen Sistem Informasi/Pusat Algoritma R&D, Universitas Minho, 4800-058 Guimar˜aes, Portugal, pcortez@dsi.uminho.pt
Halaman beranda WWW: http://www.dsi.uminho.pt/˜pcortez

Abstrak. Kebakaran hutan merupakan masalah lingkungan yang besar, yang menimbulkan kerusakan ekonomi dan ekologis sekaligus membahayakan nyawa manusia. Deteksi cepat merupakan elemen kunci untuk mengendalikan fenomena tersebut. Untuk mencapainya, salah satu alternatifnya adalah menggunakan alat otomatis berdasarkan sensor lokal, seperti yang disediakan oleh stasiun meteorologi. Kondisi meteorologi (misalnya suhu, angin) diketahui memengaruhi kebakaran hutan dan beberapa indeks kebakaran, seperti Indeks Cuaca Kebakaran Hutan (FWI), menggunakan data tersebut. Dalam karya ini, kami mengeksplorasi pendekatan Data Mining (DM) untuk memprediksi area kebakaran hutan yang terbakar. Lima teknik DM yang berbeda, misalnya Support Vector Machines (SVM) dan Random Forests, dan empat pengaturan pemilihan fitur yang berbeda (menggunakan komponen spasial, temporal, FWI, dan atribut cuaca), diuji pada data dunia nyata terkini yang dikumpulkan dari wilayah timur laut Portugal. Konfigurasi terbaik menggunakan SVM dan empat masukan meteorologi (yaitu suhu, kelembaban relatif, hujan, dan angin) dan mampu memprediksi area kebakaran pada kebakaran kecil, yang lebih sering terjadi. Pengetahuan tersebut sangat berguna untuk meningkatkan manajemen sumber daya pemadam kebakaran (misalnya memprioritaskan target untuk tanker udara dan kru darat).
Kata Kunci: Aplikasi Penambangan Data, Ilmu Kebakaran, Regresi, Mesin Vektor Pendukung.

1 Pendahuluan

Salah satu masalah lingkungan yang utama adalah terjadinya kebakaran hutan (juga disebut kebakaran hutan liar), yang mempengaruhi pelestarian hutan, menimbulkan kerusakan ekonomi dan ekologi, serta menyebabkan penderitaan manusia. Fenomena tersebut disebabkan oleh berbagai penyebab (misalnya kelalaian manusia dan petir) dan meskipun pengeluaran negara untuk mengendalikan bencana ini meningkat, setiap tahun jutaan hektar hutan **(ha)** hancur di seluruh dunia. Secara khusus, Portugal sangat terpengaruh oleh kebakaran hutan [7]. Dari tahun 1980 hingga 2005, lebih dari 2,7 juta **ha** kawasan hutan (setara dengan luas daratan Albania) telah hancur. Musim kebakaran tahun 2003 dan 2005 sangat dramatis, mempengaruhi 4,6% dan 3,1% wilayah, dengan 21 dan 18 kematian manusia.

Deteksi cepat merupakan elemen kunci untuk pemadaman kebakaran yang sukses. Karena pengawasan manusia tradisional mahal dan dipengaruhi oleh faktor subjektif, ada penekanan untuk mengembangkan solusi otomatis. Ini dapat dikelompokkan menjadi tiga kategori utama [1]: berbasis satelit, pemindai inframerah/asap dan sensor lokal (misalnya meteorologi). Satelit memiliki biaya akuisisi, penundaan lokalisasi dan resolusinya tidak memadai untuk

semua kasus. Selain itu, pemindai memiliki biaya peralatan dan perawatan yang tinggi. Kondisi cuaca, seperti suhu dan kelembaban udara, diketahui memengaruhi terjadinya kebakaran [15]. Karena stasiun meteorologi otomatis sering tersedia (misalnya Portugal memiliki 162 stasiun resmi), data tersebut dapat dikumpulkan secara real-time, dengan biaya rendah.

Di masa lalu, data meteorologi telah dimasukkan ke dalam indeks numerik, yang digunakan untuk pencegahan (misalnya memperingatkan masyarakat tentang bahaya kebakaran) dan untuk mendukung keputusan manajemen kebakaran (misalnya tingkat kesiapan, memprioritaskan target atau mengevaluasi pedoman untuk pemadaman kebakaran yang aman). Secara khusus, sistem Indeks Cuaca Kebakaran Hutan Kanada (FWI) [24] dirancang pada tahun 1970-an ketika komputer langka, sehingga hanya memerlukan perhitungan sederhana menggunakan tabel pencarian dengan bacaan dari empat pengamatan meteorologi (yaitu suhu, kelembaban relatif, hujan dan angin) yang dapat dikumpulkan secara manual di stasiun cuaca. Meskipun demikian, saat ini indeks ini sangat digunakan tidak hanya di Kanada tetapi juga di beberapa negara di seluruh dunia (misalnya Argentina atau Selandia Baru). Meskipun iklim Mediterania berbeda dengan iklim di Kanada, sistem FWI berkorelasi dengan aktivitas kebakaran di negara-negara Eropa Selatan, termasuk Portugal [26].

Di sisi lain, minat terhadap Data Mining (DM), juga dikenal sebagai Knowledge Discovery in Databases (KDD), muncul karena kemajuan Teknologi Informasi, yang menyebabkan pertumbuhan eksponensial dalam database bisnis, ilmiah, dan teknik [8].
Semua data ini menyimpan informasi berharga, seperti tren dan pola, yang dapat digunakan untuk meningkatkan pengambilan keputusan. Namun, para ahli manusia terbatas dan mungkin mengabaikan detail penting. Selain itu, analisis statistik klasik tidak dapat digunakan ketika data yang sangat banyak dan/atau kompleks tersebut tersedia. Oleh karena itu, alternatifnya adalah menggunakan alat DM otomatis untuk menganalisis data mentah dan mengekstrak informasi tingkat tinggi bagi pembuat keputusan [10].

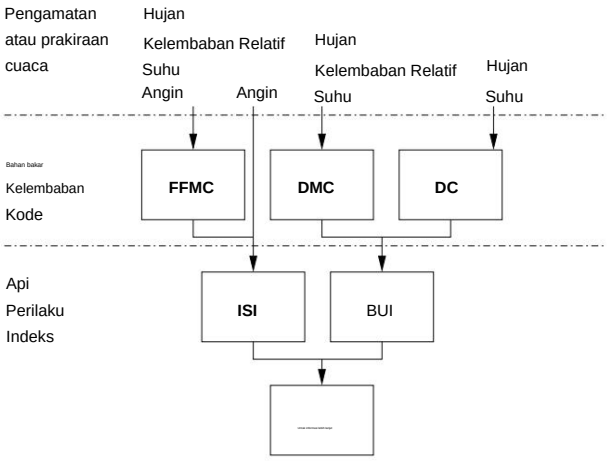
Memang, beberapa teknik DM telah diterapkan pada domain deteksi kebakaran. Misalnya, Vega-Garcia et al. [25] mengadopsi Jaringan Syaraf Tiruan (NN) untuk memprediksi terjadinya kebakaran hutan yang disebabkan manusia. Pemindai inframerah dan NN digabungkan dalam [1] untuk mengurangi alarm palsu kebakaran hutan dengan keberhasilan 90%. Pengelompokan spasial (FASTCiD) diadopsi oleh Hsu et al. [14] untuk mendeteksi titik kebakaran hutan pada citra satelit. Pada tahun 2005 [19], citra satelit dari kebakaran hutan Amerika Utara dimasukkan ke dalam Support Vector Machine (SVM), yang memperoleh akurasi 75% dalam menemukan asap pada tingkat piksel 1,1 km. Stojanova et al. [23] telah menerapkan Regresi Logistik, Hutan Acak (RF) dan Pohon Keputusan (DT) untuk mendeteksi terjadinya kebakaran di hutan Slovenia, menggunakan data berbasis satelit dan meteorologi. Model terbaik diperoleh dengan bagging DT, dengan akurasi keseluruhan 80%.

Berbeda dengan karya-karya sebelumnya, kami menyajikan pendekatan kebakaran hutan DM yang baru, yang menekankan penggunaan data meteorologi waktu nyata dan tidak mahal. Kami akan menggunakan data dunia nyata terkini, yang dikumpulkan dari wilayah timur laut Portugal, dengan tujuan untuk memperkirakan area (atau ukuran) kebakaran hutan yang terbakar. Beberapa percobaan dilakukan dengan mempertimbangkan lima teknik DM (yaitu regresi berganda, DT, RF, NN, dan SVM) dan empat pengaturan pemilihan fitur (yaitu menggunakan spasial, temporal, sistem FWI, dan data meteorologi). Solusi yang diusulkan hanya mencakup empat variabel cuaca (yaitu hujan, angin, suhu, dan kelembapan) yang dipadukan dengan SVM dan mampu memperkirakan area kebakaran pada kebakaran kecil, yang merupakan mayoritas kejadian kebakaran. Pengetahuan tersebut sangat berguna untuk mendukung keputusan manajemen kebakaran (misalnya perencanaan sumber daya).

Makalah ini disusun sebagai berikut. Pertama, kami memaparkan data kebakaran hutan di Bagian 2. Metode DM yang diadopsi disajikan di Bagian 3, sedangkan hasilnya ditunjukkan dan dibahas di Bagian 4. Terakhir, simpulan penutup ditarik (Bagian 5).

2 Data Kebakaran Hutan

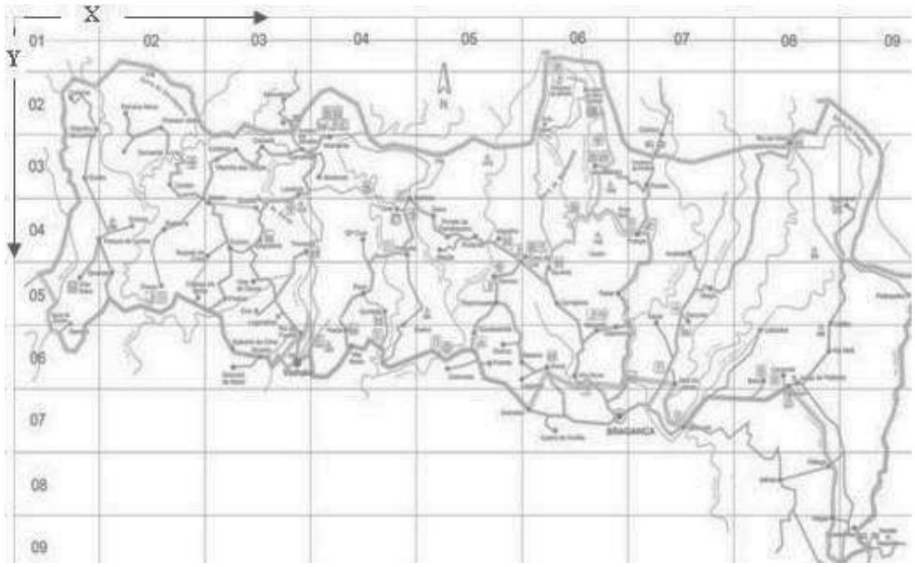
Forest Fire Weather Index (FWI) adalah sistem Kanada untuk menilai bahaya kebakaran dan mencakup enam komponen (Gambar 1) [24]: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) dan FWI. Tiga yang pertama terkait dengan kode bahan bakar: FFMC menunjukkan kadar air serasah permukaan dan memengaruhi penyalaan dan penyebaran api, sedangkan DMC dan DC menunjukkan kadar air lapisan organik dangkal dan dalam, yang memengaruhi intensitas api. ISI adalah skor yang berkorelasi dengan kecepatan penyebaran api, sedangkan BUI menunjukkan jumlah bahan bakar yang tersedia. Indeks FWI adalah indikator intensitas api dan menggabungkan dua komponen sebelumnya. Meskipun skala yang berbeda digunakan untuk setiap elemen FWI, nilai yang tinggi menunjukkan kondisi pembakaran yang lebih parah. Selain itu, kode kelembaban bahan bakar memerlukan memori (jeda waktu) dari kondisi cuaca masa lalu: 16 jam untuk FFMC, 12 hari untuk DMC dan 52 hari untuk DC.



Gambar 1. Struktur Indeks Cuaca Kebakaran (diadaptasi dari [24])

Studi ini akan mempertimbangkan data kebakaran hutan dari taman alam Montesinho, dari wilayah timur laut Tr´as-os-Montes di Portugal (Gambar 2). Taman ini memiliki keanekaragaman flora dan fauna yang tinggi. Berada dalam iklim supra-Mediterania, suhu tahunan rata-rata berada dalam kisaran 8 hingga 12ÿC. Data yang digunakan dalam percobaan dikumpulkan dari Januari 2000 hingga Desember 2003 dan dibuat menggunakan dua sumber. Basis data pertama dikumpulkan oleh inspektur yang bertanggung jawab atas kejadian kebakaran Montesinho. Setiap hari, setiap kali kebakaran hutan terjadi, beberapa fitur

telah terdaftar, seperti waktu, tanggal, lokasi spasial dalam grid 9×9 **(sumbu** x dan **y)**
Gambar 2), jenis vegetasi yang terlibat, enam komponen sistem FWI
dan total area yang terbakar. Basis data kedua dikumpulkan oleh Institut Politeknik Braganc¸a, yang berisi
beberapa pengamatan cuaca (misalnya kecepatan angin) yang
direkam dengan periode 30 menit oleh stasiun meteorologi yang terletak di pusat
taman Montesinho. Kedua basis data tersebut disimpan dalam puluhan lembar kerja terpisah, dalam
format yang berbeda, dan upaya manual yang substansial dilakukan untuk mengintegrasikannya ke dalam
satu set data dengan total 517 entri. Data ini tersedia di:
http://www.dsi.uminho.pt/˜pcortez/kebakaran hutan/.



Gambar 2. Peta taman alam Montesinho

Tabel 1 menunjukkan deskripsi fitur data yang dipilih. Empat baris pertama menunjukkan
atribut spasial dan temporal. Hanya dua fitur geografis yang disertakan, yaitu
Nilai sumbu **X** dan **Y** di mana kebakaran terjadi, karena jenis vegetasi yang ditampilkan
kualitas rendah (yaitu lebih dari 80% nilai hilang). Setelah berkonsultasi dengan inspektur kebakaran Montesinho, kami memilih variabel temporal **bulan** dan **hari** dalam seminggu.
Kondisi cuaca bulanan rata-rata cukup berbeda, sedangkan hari dalam seminggu bisa
juga mempengaruhi kebakaran hutan (misalnya hari kerja vs hari libur) karena sebagian besar kebakaran disebabkan oleh manusia
penyebabnya. Berikutnya adalah empat komponen FWI yang secara langsung dipengaruhi oleh cuaca
kondisi cuaca (Gambar 1, dicetak tebal). BUI dan FWI dibuang karena keduanya bergantung pada nilai
sebelumnya. Dari database stasiun meteorologi, kami memilih
empat atribut cuaca yang digunakan oleh sistem FWI. Berbeda dengan jeda waktu yang digunakan oleh
FWI, dalam kasus ini nilai-nilai menunjukkan catatan instan, seperti yang diberikan oleh sensor stasiun saat
kebakaran terdeteksi. Pengecualiannya adalah variabel **hujan ,** yang menunjukkan akumulasi
curah hujan dalam 30 menit sebelumnya.

Area yang terbakar ditunjukkan pada Gambar 3, yang menunjukkan kemiringan positif, dengan mayoritas kebakaran menunjukkan ukuran kecil. Perlu dicatat bahwa sifat miring ini juga ada di negara lain, seperti Kanada [18]. Mengenai kumpulan data saat ini, ada 247 sampel dengan nilai nol. Seperti yang dinyatakan sebelumnya, semua entri menunjukkan kejadian kebakaran dan nilai nol berarti bahwa area yang terbakar kurang dari 1ha/100 = 100m2 . Untuk mengurangi kemiringan dan meningkatkan simetri, fungsi logaritma **y** = ln(x + 1), yang merupakan transformasi umum yang cenderung meningkatkan hasil regresi untuk target miring ke kanan [20], diterapkan pada atribut **area** (Gambar 3). Variabel transformasi akhir akan menjadi target keluaran dari pekerjaan ini.

Tabel 1. Atribut dataset yang telah diproses sebelumnya

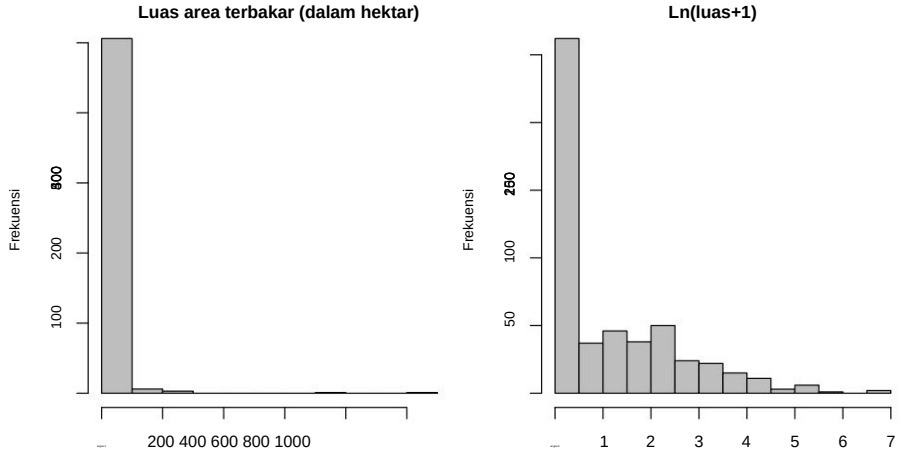
Deskripsi Atribut	
Koordinat sumbu **x** (dari 1 sampai 9) Koordinat sumbu	
kamu	y (dari 1 sampai 9) **bulan Bulan** dalam
tahun (Januari sampai Desember) **hari**	
	Hari dalam seminggu (Senin sampai Minggu)
Kode **FFMC**	
Kode **DMC** DMC	
DC	Kode DC
ISI	indeks ISI
temp Suhu luar (dalam ÿC)	
RH Kelembaban relatif luar ruangan (dalam %)	
jam) **angin**	Kecepatan angin luar (dalam km/
hujan	Curah hujan di luar ruangan (dalam mm/m2)
daerah	Total area terbakar (dalam **ha)**

3 Model Penambangan Data

Kumpulan data regresi **D** terdiri dari **k** ÿ {1, ..., N} contoh, yang masing-masing memetakan input **xk** vektor **(x** **1 buah** A) ke target **yk yang diberikan.** Kesalahan diberikan oleh: **ek** = **yk** ÿ **yk,** di mana **yk** mewakili nilai prediksi untuk pola input **k .** Kinerja keseluruhan dihitung dengan metrik global, yaitu *Mean Absolute Deviation (MAD)* dan *Root Mean Squared (RMSE),* yang dapat dihitung sebagai [27]:

$$\textbf{MAD} = \textbf{1/N} \times \sum_{\textbf{Ni=1}} |\textbf{yi} \ddot{y} \textbf{yi}| \qquad \textbf{Nilai RMSE} = \sqrt{\sum_{\textbf{Ni=1}} (\textbf{yi} \ddot{y} \textbf{yi})^2 / \textbf{N}} \qquad (1)$$

Pada kedua metrik tersebut, nilai yang lebih rendah menghasilkan model prediksi yang lebih baik. Namun, **RMSE** lebih sensitif terhadap kesalahan yang tinggi. Kemungkinan lain untuk membandingkan model regresi adalah kurva Regression Error Characteristic (REC) [2], yang memplot toleransi kesalahan **(sumbu** x), diberikan dalam bentuk deviasi absolut, versus persentase poin yang diprediksi.



Gambar 3. Histogram untuk area yang terbakar (kiri) dan transformasi logaritma masing-masing (kanan)

dalam toleransi (sumbu y). Regresor ideal harus menyajikan area REC yang dekat dengan

Beberapa algoritma DM, masing-masing dengan tujuan dan kemampuannya sendiri, telah diusulkan untuk tugas regresi. Pekerjaan ini akan mempertimbangkan lima model DM. Model Regresi Berganda (MR) mudah ditafsirkan dan pendekatan klasik ini telah yang banyak digunakan [11]. Namun, ia hanya dapat mempelajari pemetaan linier. Untuk mengatasi kelemahan ini, salah satu alternatifnya adalah dengan menggunakan metode yang berbasis pada struktur pohon, seperti Pohon Keputusan (Decision Trees/DT) dan Hutan Acak (RF), atau fungsi nonlinier, seperti Jaringan Syaraf (NN) dan Mesin Vektor Pendukung (SVM).

DT adalah struktur percabangan yang mewakili serangkaian aturan, yang membedakan nilai-nilai dalam bentuk hirarkis [4]. Representasi ini dapat diterjemahkan ke dalam sekumpulan aturan IF-THEN, yang mudah dipahami oleh manusia. RF [3] merupakan ensembel **T** yang tidak dipangkas DT, menggunakan pemilihan fitur acak dari sampel pelatihan bootstrap. Prediktor RF dibangun dengan merata-ratakan keluaran pohon **T.** Secara umum, RF menunjukkan signifikansi peningkatan dibandingkan dengan DT tunggal.

NN adalah model koneksionis yang terinspirasi oleh perilaku otak manusia. Secara khusus, multilayer perceptron adalah arsitektur NN yang paling populer. Ini terdiri dari jaringan umpan maju di mana neuron pemrosesan dikelompokkan menjadi beberapa lapisan dan dihubungkan dengan link tertimbang [12]. Penelitian ini akan mempertimbangkan multilayer perceptrons dengan satu hidden lapisan **H** node tersembunyi dan fungsi aktivasi logistik dan satu node keluaran dengan fungsi linier [11]. Karena fungsi biaya NN bersifat nonkonveks (dengan beberapa minimum), **NR** berjalan akan diterapkan pada setiap konfigurasi saraf, dipilih NN dengan kesalahan yang dihukum terendah. Dengan pengaturan ini, kinerja NN akan bergantung pada nilai dari H.

SVM memberikan keuntungan teoritis dibandingkan NN, seperti tidak adanya minimum lokal dalam fase optimasi model. Dalam regresi SVM, input **x** ÿ **ÿA** ditransformasikan ke dalam ruang fitur m-dimensi tinggi , dengan menggunakan pemetaan nonlinier. Kemudian, *SVM*

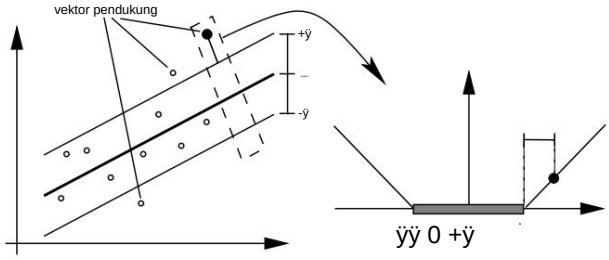
menemukan hiperbidang pemisah linier terbaik dalam ruang fitur:

$$\textbf{kamu} = \textbf{w0} + \sum_{saya=1}^{M} \textbf{dengan} ÿi(x) \quad (2)$$

di mana ÿi(x) merupakan representasi dari transformasi nonlinier, sesuai dengan fungsi kernel **K(x,** xÿ) = ÿi(x)ÿi(x Fungsi Basis Radial $\sum_{saya=1}^{M}$ '). Untuk memperkirakan SVM terbaik, fungsi kerugian tidak sensitif ÿ- (Gambar 4) sering digunakan [22]. Kernel Fungsi Basis Radial yang populer, yang menyajikan lebih sedikit hiperparameter dan kesulitan numerik dibandingkan kernel lainnya (misalnya polinomial atau sigmoid), juga akan diadopsi [13]:

$$\textbf{Rumus K(x,} \text{ xÿ) = exp(ÿÿ||x ÿ } \textbf{x} \text{ ÿ ||2), } \textbf{ÿ >} 0 \quad (3)$$

Performa SVM dipengaruhi oleh tiga parameter: **C** – trade-off antara kompleksitas model dan jumlah toleransi penyimpangan lebih besar dari **ÿ ; ÿ** – lebar zona tidak sensitif ÿ; dan **ÿ** – parameter kernel. Karena ruang pencarian untuk ketiga parameter tersebut tinggi, nilai **C** dan **ÿ** akan ditetapkan menggunakan heuristik yang diusulkan dalam [5]: **C** = 3 (untuk input standar) dan **ÿ** = **3ÿ** dan **ÿ** adalah $\frac{\text{dalam(N)}}{\text{N}}$, Di mana deviasi standar seperti yang diprediksi oleh algoritma 3 tetangga terdekat.



Gambar 4. Contoh regresi SVM linier dan fungsi kerugian tidak sensitif ÿ (diadaptasi dari [22])

Karena kinerjanya dalam hal pengetahuan prediktif, RF, NN, dan SVM mendapatkan perhatian dalam bidang DM [27]. Namun, metode ini memerlukan lebih banyak komputasi dan menggunakan representasi yang lebih sulit untuk ditafsirkan jika dibandingkan dengan model MR dan DT yang lebih sederhana. Meskipun demikian, masih mungkin untuk memberikan pengetahuan penjelasan untuk RF, NN, dan SVM dalam hal relevansi input [3][16].

4 Hasil Eksperimen

Semua percobaan yang dilaporkan dalam penelitian ini dilakukan menggunakan **RMiner** [6], pustaka sumber terbuka untuk lingkungan statistik **R** [21] yang memfasilitasi penggunaan teknik DM dalam tugas klasifikasi dan regresi. Secara khusus, **RMiner** menggunakan **paket randomForest** (algoritma RF oleh L. Breiman dan A. Cutler), **nnet** (untuk NN) dan **kernlab** (alat LIBSVM [13]).

Sebelum memasang model, beberapa praproses diperlukan oleh model MR, NN, dan SVM. Variabel nominal (yaitu diskrit dengan lebih dari dua nilai tidak berurutan), seperti **bulan** dan **hari,** diubah menjadi pengodean *1-dari-C ,* seperti yang disarankan dalam [13]. Selain itu, untuk metode NN dan SVM, semua atribut distandarisasi ke mean nol dan satu deviasi standar [11]. Berikutnya, model regresi dipasang. Parameter MR dioptimalkan menggunakan algoritma kuadrat terkecil, sementara pemisahan simpul DT disesuaikan untuk pengurangan jumlah kuadrat. Mengenai metode yang tersisa, parameter default diadopsi untuk RF (misalnya **T** = 500), NN disesuaikan menggunakan **NR** = 3 pelatihan dan **E** = 100 epoch dari algoritma BFGS dan algoritma Optimasi Minimal Berurutan digunakan untuk memasang SVM. Setelah memasang model DM, output diproses pasca menggunakan kebalikan dari transformasi logaritma. Dalam beberapa kasus, transformasi ini dapat menghasilkan angka negatif dan output negatif tersebut ditetapkan ke

nol.

Untuk menyimpulkan tentang dampak variabel input, empat pengaturan pemilihan fitur berbeda diuji untuk setiap algoritma DM: **STFWI** – menggunakan spasial, temporal dan empat komponen FWI; **STM** – dengan spasial, temporal dan empat variabel cuaca; **FWI** – hanya menggunakan empat komponen FWI; dan **M** – dengan empat kondisi cuaca. Untuk mengakses kinerja prediktif, tiga puluh kali menjalankan 10-fold [17] (total 300 simulasi) diterapkan pada setiap konfigurasi yang diuji. Mengenai hiperparameter NN dan SVM, pencarian grid internal 10-fold (yaitu hanya menggunakan data pelatihan) digunakan untuk menemukan 2 ÿ1} terbaik. Setelah memilih nilai H/ÿ , **H** ÿ {2, **4, 6, 8,** 10} dan **ÿ** ÿ {2 model NN/SVM dilatih ulang dengan semua data pelatihan. Tabel 2 menunjukkan nilai median dari parameter **H** dan **ÿ** yang dipilih .

-9 **-7** 2 2 **-5** 2 **-3**

Tabel 2. Hiperparameter terbaik untuk NN dan SVM (nilai median)

	Pengaturan Pemilihan Fitur			
Model DM	STFWI	STM	FWI	M
NN	6	4	4	4
SVM	**2ÿ5** 2	-3	2 -3	2 -3

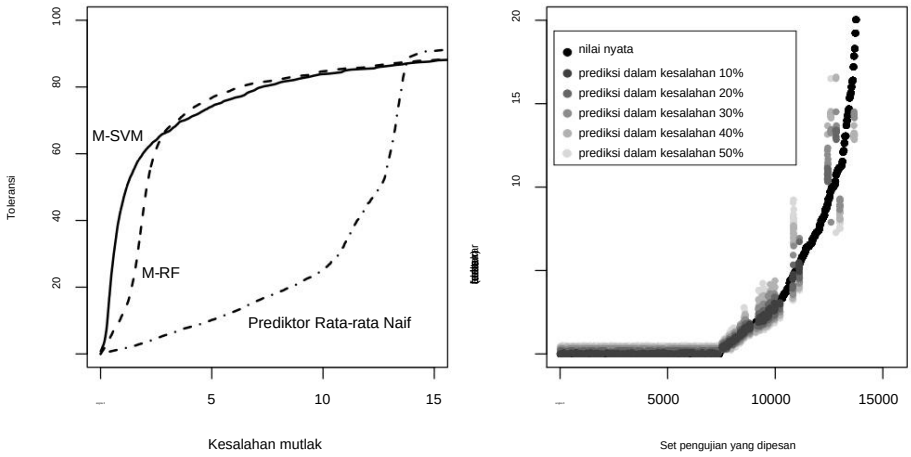
Hasilnya ditunjukkan pada Tabel 3 dalam bentuk rata-rata dan interval kepercayaan t-student 95% masing-masing [9]. Untuk tujuan pembandingan, prediktor rata-rata naif (baris pertama) juga ditambahkan ke tabel. Di bawah kriteria **MAD ,** semua metode DM mengungguli tolok ukur naif. Dalam pemilihan fitur tertentu, SVM cenderung menghasilkan prediksi terbaik (kecuali untuk pengaturan STM). Hasil menarik lainnya adalah tidak relevannya variabel spasial dan temporal, karena ketika dihilangkan, kinerja SVM meningkat. Akibatnya, konfigurasi terbaik diberikan oleh pengaturan **M** dan model SVM dan uji-t berpasangan terhadap semua model lain mengonfirmasi signifikansi statistik dari hasil ini. Untuk SVM, lebih baik menggunakan kondisi cuaca daripada variabel FWI. Ini adalah hasil yang menarik, karena variabel meteorologi dapat diperoleh langsung dari sensor cuaca, tanpa perlu perhitungan akumulasi. Namun, dari sudut pandang **RMSE ,** opsi terbaik adalah prediktor rata-rata naif.

kontradiksi induk dibenarkan oleh sifat masing-masing kriteria kesalahan, yaitu **RMSE** adalah lebih sensitif terhadap outlier dibandingkan metrik **MAD .**

Analisis yang lebih rinci terhadap kualitas kesalahan prediksi diberikan dengan menggunakan REC kurva (Gambar 5). Untuk menyederhanakan visualisasi, hanya tiga model yang diplot: M–SVM, konfigurasi **MAD** terbaik ; M–RF, metode berbasis meteorologi terbaik kedua (dalam dalam hal nilai **MAD);** dan Naive, model **RMSE** terbaik . Dari analisis REC, M–SVM jelas merupakan solusi terbaik, dengan area tertinggi. Meskipun hanya ada 0,22 perbedaan dalam hal nilai **MAD** rata-rata , kurva M–SVM dan M–RF berbeda, dengan model sebelumnya menyajikan prediksi terbaik untuk probabilitas yang dapat diterima. kesalahan absolut hingga 2,85. Misalnya, 46% contoh diprediksi secara akurat jika kesalahan **1ha** diterima dan nilai ini meningkat menjadi 61% ketika kesalahan yang dapat diterima kesalahannya **2ha.** Mengenai prediktor naif, ini adalah metode terburuk, melampaui yang lain alternatif hanya setelah kesalahan absolut 13,7.

Tabel 3. Hasil prediksi dalam bentuk kesalahan **MAD (nilai RMSE** dalam tanda kurung; garis bawah – model terbaik; **tebal** – terbaik dalam pemilihan fitur)

DM	**Pengaturan Pemilihan Fitur**			
Model STFWI	STM		M	
Naif 18,61±0,01 (63,7±0,0)	18,61±0,01 (63,7±0,0)	18,61±0,01 (63,7±0,0)	18,61±0,01 (63,7±0,0)	
MR 13,07±0,01 (64,5±0,0)	13,04±0,01 (64,4±0,0)	13,00±0,00 (64,5±0,0)	13,01±0,00 (64,5±0,0)	
Tanggal Lahir 13,46±0,04 (64,4±0,1)	13,43±0,06 (64,6±0,0)	13,24±0,03 (64,4±0,0)	13,18±0,05 (64,5±0,0)	
Tingkat bunga 13,31±0,02 (64,3±0,0)	13,04±0,01 (64,5±0,0)	13,38±0,05 (64,0±0,1)	12,93±0,01 (64,4±0,0)	
NNN 13,09±0,04 (64,5±0,0)	13,92±0,60 (68,9±8,5)	13,08±0,05 (64,6±0,1)	13,71±0,69 (66,9±3,4)	
SVM 13,07±0,04 (64,7±0,0)	13,13±0,02 (64,7±0,0)	12,86±0,00 (64,7±0,0)	12,71±0,01 (64,7±0,0)	



Gambar 5. Kurva REC untuk model M-SVM, M-RF dan Naive (kiri); dan nilai riil (hitam) titik-titik) dan prediksi M-SVM (titik-titik abu-abu) sepanjang rentang keluaran sumbu y (kanan)

Untuk melengkapi analisis REC, plot lain disajikan untuk konfigurasi M–SVM (Gambar 5). Tujuannya adalah untuk mengamati bagaimana kesalahan didistribusikan sepanjang rentang keluaran. Nilai riil (titik hitam) dari set pengujian diurutkan (sumbu x) sesuai dengan luas area yang terbakar (sumbu y). Perlu dicatat bahwa sumbu x berkisar dari 1 hingga 517×30 kali percobaan = 15510. Untuk memperjelas analisis, sumbu y ditetapkan dalam rentang [0, **20 ha].** Prediksi M–SVM juga ditunjukkan dalam gambar, menggunakan skala abu-abu yang bergantung pada akurasi. Secara umum, titik abu-abu menunjukkan prediksi dalam kesalahan relatif yang berkisar dari 10% (abu-abu lebih gelap) hingga 50% (abu-abu lebih terang). Pengecualiannya adalah ketika nilai riil di bawah **1 ha.** Dalam kasus ini, skala abu-abu sesuai dengan perbedaan absolut (dari 0,1 ha hingga **0,5 ha).** Grafik tersebut menunjukkan bahwa kinerja M–SVM lebih baik ketika memprediksi kebakaran kecil (misalnya dalam kisaran [0, **3,2 ha]).**

Mengenai prosedur relevansi input, seluruh 517 record digunakan untuk menyesuaikan model M–SVM. Kemudian, prosedur analisis sensitivitas [16] dilakukan dengan mengukur varians **(Va)** yang dihasilkan oleh output ketika atribut input **xa** yang diberikan bervariasi melalui seluruh rentangnya dengan level **L** (di sini ditetapkan ke **L** = 5). Biarkan **yaLi** menjadi output rata-rata ketika atribut **xa** = **Li** dan semua input lainnya ditetapkan ke nilai aslinya **i=1 (yaLi** ÿ **yaLi)** 2/(L ÿ 1). Varians ini dapat (dari kumpulan data). Kemudian **Va** 4). Prosedur ini menunjukkan bahwa = direlatifkan, dengan menggunakan ekspresi: **Ra** = **Va/ Vj** (Tabel semua kondisi cuaca memengaruhi model, dengan suhu luar menjadi fitur yang paling penting, diikuti oleh akumulasi presipitasi (hujan).

Tabel 4. Nilai analisis sensitivitas untuk masukan cuaca model M–SVM

suhu RH angin hujan
Va 9,95 0,56 0,64 2,45 Ra
73,2% 4,1% 4,7% 18,0%

5 Kesimpulan

Kebakaran hutan menyebabkan kerusakan lingkungan yang signifikan dan mengancam kehidupan manusia. Dalam dua dekade terakhir, upaya substansial telah dilakukan untuk membangun alat deteksi otomatis yang dapat membantu Sistem Manajemen Kebakaran (FMS). Tiga tren utama adalah penggunaan data satelit, pemindai inframerah/asap, dan sensor lokal (misalnya meteorologi). Dalam karya ini, kami mengusulkan pendekatan Penambangan Data (DM) yang menggunakan data meteorologi, sebagaimana yang dideteksi oleh sensor lokal di stasiun cuaca, dan yang diketahui memengaruhi kebakaran hutan.
Keuntungannya adalah data tersebut dapat dikumpulkan secara real-time dan dengan biaya yang sangat rendah, jika dibandingkan dengan pendekatan satelit dan pemindai. Data dunia nyata terkini, dari wilayah timur laut Portugal, digunakan dalam percobaan. Basis data tersebut mencakup komponen spasial, temporal, dari Indeks Cuaca Kebakaran Kanada (FWI) dan empat kondisi cuaca. Masalah ini dimodelkan sebagai tugas regresi, yang tujuannya adalah prediksi area yang terbakar. Lima algoritme DM yang berbeda, termasuk Dukungan

Mesin Vektor (SVM), dan empat pilihan fitur (menggunakan kombinasi berbeda dari elemen spasial, temporal, FWI, dan variabel meteorologi) diuji.

Solusi yang diusulkan, yang berbasis pada SVM dan hanya memerlukan empat masukan cuaca langsung (yaitu suhu, hujan, kelembaban relatif, dan kecepatan angin) mampu memprediksi kebakaran kecil, yang merupakan mayoritas kejadian kebakaran. Kelemahannya adalah akurasi prediksi yang lebih rendah untuk kebakaran besar. Sepengetahuan kami, ini adalah pertama kalinya area kebakaran diprediksi hanya menggunakan data berbasis meteorologi dan diperlukan penelitian eksplorasi lebih lanjut. Seperti yang dikemukakan dalam [18], memprediksi ukuran kebakaran hutan merupakan tugas yang menantang. Untuk memperbaikinya, kami yakin bahwa informasi tambahan (tidak tersedia dalam studi ini) diperlukan, seperti jenis vegetasi dan intervensi pemadaman kebakaran (misalnya waktu yang telah berlalu dan strategi pemadaman kebakaran). Meskipun demikian, model yang diusulkan masih berguna untuk meningkatkan manajemen sumber daya pemadaman kebakaran. Misalnya, ketika kebakaran kecil diprediksi maka tanker udara dapat dihemat dan kru darat kecil dapat dikirim. Manajemen semacam itu akan sangat menguntungkan di musim kebakaran yang dramatis, ketika kebakaran simultan terjadi di lokasi yang berbeda.

Studi ini didasarkan pada pembelajaran off-line, karena teknik DM diterapkan setelah data dikumpulkan. Namun, pekerjaan ini membuka ruang untuk pengembangan alat otomatis untuk dukungan manajemen kebakaran. Memang, di masa depan kami bermaksud untuk menguji pendekatan yang diusulkan dengan menggunakan lingkungan pembelajaran on-line sebagai bagian dari FMS. Ini akan memungkinkan kami untuk mendapatkan umpan balik yang berharga dari manajer pemadam kebakaran, dalam hal kepercayaan dan penerimaan solusi alternatif ini. Kemungkinan menarik lainnya adalah penggunaan prakiraan cuaca, untuk membangun respons proaktif. Karena sistem FWI digunakan secara luas di seluruh dunia, penelitian lebih lanjut diperlukan untuk mengonfirmasi apakah kondisi cuaca langsung lebih disukai daripada nilai akumulasi, seperti yang disarankan oleh studi ini. Akhirnya, karena kebakaran besar merupakan kejadian langka, teknik deteksi outlier [28] juga akan dibahas.

6 Ucapan Terima Kasih

Kami ingin mengucapkan terima kasih kepada Manuel Rainha atas penyediaan data spasial, temporal, dan FWI. Kami juga berterima kasih kepada Institut Politeknik Braganc¸a atas basis data stasiun meteorologi.

Referensi

1. B. Arrue, A. Ollero, dan J. Matinez de Dios. Sistem Cerdas untuk Mengurangi Alarm Palsu dalam Deteksi Kebakaran Hutan Inframerah. *Sistem Cerdas IEEE,* 15(3):64–73, 2000.
2. J. Bi dan K. Bennett. Kurva Karakteristik Kesalahan Regresi. Dalam *Prosiding Konferensi Internasional ke-20 tentang Pembelajaran Mesin (ICML),* halaman 43–50, Washington DC, AS, 2003.
3. L. Breiman. Hutan Acak. *Pembelajaran Mesin,* 45(1):5–32, 2001.
4. L. Breiman, J. Friedman, R. Ohlsen, dan C. Stone. *Pohon Klasifikasi dan Regresi.* Wadsworth, Monterey, CA, 1984.
5. V. Cherkassy dan Y. Ma. Pemilihan Praktis Parameter SVM dan Estimasi Noise untuk Regresi SVM. *Jaringan Syaraf Tiruan,* 17(1):113–126, 2004.
6. P. Cortez. RMiner: Penambangan Data dengan Jaringan Syaraf Tiruan dan Mesin Vektor Pendukung menggunakan R. Dalam R. Rajesh (Ed.), *Pendahuluan tentang Perangkat Lunak dan Kotak Peralatan Ilmiah Tingkat Lanjut,* Siap Diterbitkan.

7. Komisi Eropa. Kebakaran Hutan di Eropa. Laporan teknis, Laporan N-4/6, 2003/2005.
8. U. Fayyad, G. Piatetsky-Shapiro, dan P. Smyth. *Kemajuan dalam Penemuan Pengetahuan dan Penambangan Data.* MIT Press, 1996.
9. A. Flexer. Evaluasi statistik eksperimen jaringan saraf: Persyaratan minimum dan praktik terkini. Dalam *Prosiding Pertemuan Eropa ke-13 tentang Sibernetika dan Penelitian Sistem,* volume 2, halaman 1005–1008, Wina, Austria, 1996.
10. D. Hand, H. Mannila, dan P. Smyth. *Prinsip Penambangan Data.* MIT Press, Cambridge, MA, 2001.
11. T. Hastie, R. Tibshirani, dan J. Friedman. *Elemen Pembelajaran Statistik: Penambangan Data, Inferensi, dan Prediksi.* Springer-Verlag, NY, AS, 2001.
12. S. Haykin. *Jaringan Syaraf Tiruan - Landasan Komprehensif.* Prentice-Hall, New Jersey, 2nd edisi, 1999.
13. C. Hsu, C. Chang, dan C. Lin. Panduan Praktis untuk Klasifikasi Vektor Pendukung. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, Juli, Dep. Ilmu Komp. dan Teknik Informasi, Universitas Nasional Taiwan, 2003.
14. W. Hsu, M. Lee, dan J. Zhang. Image Mining: Tren dan Perkembangan. *Jurnal Intelijen Sistem Informasi Cerdas,* 19(1):7–23, 2002.
15. J. Terradas J. Pinol dan F. Lloret. Pemanasan iklim, bahaya kebakaran hutan, dan kejadian kebakaran hutan di pesisir timur Spanyol. *Perubahan Iklim,* 38:345–357, 1998.
16. R. Kewley, M. Embrechts, dan C. Breneman. Penambangan Strip Data untuk Desain Virtual Farmasi dengan Jaringan Syaraf Tiruan. *Transaksi IEEE pada Jaringan Syaraf Tiruan,* 11(3):668–679, Mei 2000.
17. R. Kohavi. Sebuah Studi tentang Validasi Silang dan Bootstrap untuk Estimasi Akurasi dan Pemilihan Model. Dalam *Prosiding Konferensi Gabungan Internasional tentang Kecerdasan Buatan (IJCAI),* Montreal, Quebec, Kanada, Agustus 1995.
18. K. Malarz, S. Kaczanowska, dan K. Kulakowski. Apakah kebakaran hutan dapat diprediksi? *Internasional Jurnal Fisika Modern,* 13(8):1017–1031, 2002.
19. D. Mazzoni, L. Tong, D. Diner, Q. Li, dan J. Logan. Menggunakan Data MISR dan MODIS untuk Mendeteksi dan Menganalisis Ketinggian Asap di Amerika Utara Selama Musim Panas 2004. *Abstrak Pertemuan Musim Gugur AGU,* halaman B853+, Desember 2005.
20. S. Menard. *Analisis Regresi Logistik Terapan.* SAGE, edisi ke-2, 2001.
21. Tim Inti Pengembangan R. *R: Bahasa dan lingkungan untuk komputasi statistik.* Yayasan R untuk Komputasi Statistik, Wina, Austria, 2006. URL: http://www.R-project.org, ISBN 3-900051-00-3.
22. A. Smola dan B. Scholkopf. Tutorial tentang regresi vektor pendukung. Laporan Teknis NC2-TR-1998-030, Universitas London, Inggris, 1998.
23. D. Stojanova, P. Panov, A. Kobler, S. Dzeroski, dan K. Taskova. Belajar Memprediksi Kebakaran Hutan dengan Teknik Penambangan Data yang Berbeda. Dalam D. Mladenic dan M. Grobelnik, editor, *Masyarakat Informasi Multikonferensi Internasional ke-9 (IS 2006),* Ljubljana, Slovenia, 2006.
24. S. Taylor dan M. Alexander. Sains, teknologi, dan faktor manusia dalam pemeringkatan bahaya kebakaran: pengalaman Kanada. *Jurnal Internasional Kebakaran Hutan Liar,* 15:121–135, 2006.
25. C. Vega-Garcia, B. Lee, P. Woodard, dan S. Titus. Penerapan teknologi jaringan saraf untuk memprediksi kejadian kebakaran hutan yang disebabkan oleh manusia. *Aplikasi AI,* 10(3):9–18, 1996.
26. D. Viegas, G. Biovio, A. Ferreira, A. Nosenzo, dan B. Sol. Studi Komparatif berbagai metode evaluasi bahaya kebakaran di Eropa Selatan. *Jurnal Internasional Kebakaran Hutan Liar,* 9:235–246, 1999.
27. IH Witten dan E. Frank. *Penambangan Data: Alat dan Teknik Pembelajaran Mesin Praktis dengan Implementasi Java.* Morgan Kaufmann, San Francisco, CA, 2005.
28. J. Zhao, C. Lu, dan Y. Kou. Mendeteksi Outlier Wilayah dalam Data Meteorologi. Dalam *Prosiding Simposium Internasional ACM ke-11 tentang Kemajuan dalam Sistem Informasi Geografis,* halaman 49–55, New Orleans, Louisiana, AS, 2003.