HASIL CEK_17 Perbandingan

by 17 Perbandingan

Submission date: 18-May-2022 09:32AM (UTC+0700)
Submission ID: 1838783561
File name: 17 Perbandingan.pdf (548.76K)
Word count: 4623
Character count: 27352

Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 2 (2020) 329 - 335 ISSN Media Elektronik: 2580-0760

Perbandingan Metode SVM, *RF* dan *SGD* untuk Penentuan Model Klasifikasi Kinerja *Programmer* pada Aktivitas Media Sosial

Rusydi Umar[1], Imam Riadi[2], Purwono[3]
[1,3]Program Studi Teknik Informatika, Universitas Ahmad Dahlan
[2]Program Studi Sistem Informasi, Universitas Ahmad Dahlan
[1]rusydi@mti.uad.ac.id, [2]imam.riadi@is.uad.ac.id, [3]purwono1907048015@webmail.uad.ac.id

Abstract

The failure of most startups in Indonesia is caused by team performance that is not solid and competent. Programmers are an integral profession in a startup team. The development of social media can be used as a strategic tool for recruiting the best programmer candidates in a company. This strategic tool is in the form of an automatic classification system of social media posting from prospective programmers. The classification results are expected to be able to predict the performance patterns of each candidate with a predicate of good or bad performance. The classification method with the best accuracy needs to be chosen in order to get an effective strategic tool so that a comparison of several methods is needed. This study compares classification methods including the Support Vector Machines (SVM) algorithm, Random Forest (RF) and Stochastic Gradient Descent (SGD). The classification results show the percentage of accuracy with k = 10 cross validation for the SVM algorithm reaches 81.3%, RF at 74.4%, and SGD at 80.1% so that the SVM method is chosen as a model of programmer performance classification on social media activities.

Keywords: classification, support vector machine, random forest, stochastic gradient descent, programmer

Abstrak

Kegagalan kebanyakan *startup* di Indonesia disebabkan oleh kinerja tim yang tidak solid dan kompeten. *Programmer* merupakan profesi yang tidak terpisahkan dalam sebuah tim *startup*. Perkembangan media sosial bisa digunakan sebagai alat strategis untuk merekrut kandidat *programmer* terbaik dalam suatu perusahaan. Alat strategis ini berupa sistem klasifikasi otomatis postingan media sosial dari kandidat *programmer*. Hasil klasifikasi diharapkan mampu memprediksi pola kinerja dari setiap kandidat dengan predikat kinerja baik atau buruk. Metode klasifikasi dengan akurasi terbaik perlu dipilih agar mendapatkan alat strategis yang efektif sehingga diperlukan adanya perbandingan dari beberapa metode. Penelitian ini melakukan perbandingan metode klasifikasi antara lain algoritma *Support Vector Machines* (SVM), *Random Forest* (RF) dan *Stochastic Gradient Descent* (SGD). Hasil klasifikasi menunjukan persentase akurasi dengan k=10 *cross validation* untuk algoritma SVM mencapai angka 81,3%, RF pada angka 74,4 %, dan SGD pada angka 80,1% sehingga metode SVM dipilih sebagai model klasifikasi kinerja *programmer* pada aktivitas media sosial.

Kata kunci: klasifikasi, *support vector machine, random forest, stochastic gradient descent, programmer*

1. Pendahuluan

Pertumbuhan ekonomi di Indonesia dipengaruhi oleh meningkatnya perkembangan teknologi. Berbagai jenis perusahaan baru berbasis teknologi informasi seperti *startup* tumbuh dan berkembang cukup pesat. Kesuksesan besar yang dialami Tokopedia, Go-Jek, Bukalapak dan Traveloka menjadi salah satu pemicu tumbuhnya *startup* baru di Indonesia. *Startup* merupakan sebuah wadah yang digunakan untuk menemukan model bisnis baru agar mendapatkan keuntungan yang besar [1]. Kesuksesan beberapa *startup* juga diimbangi dengan banyaknya kegagalan *startup* lain. Salah satu faktor kegagalan ialah tim yang bekerja tidak solid dan tidak kompeten [2].

Programmer merupakan salah satu profesi yang berpengaruh penting pada kinerja *startup*. Pemilihan kandidat dalam profesi ini harus dilakukan dengan prosedur yang benar. Peningkatan penggunaan media sosial yang begitu pesat bisa digunakan sebagai salah satu alat strategis untuk merekrut kandidat profesional

Diterima Redaksi : 13-03-2020 | Selesai Revisi : 29-03-2020 | Diterbitkan Online : 20-04-2020

dalam suatu perusahaan [3]. Media sosial juga dapat digunakan untuk melihat aspek keprofesionalan seseorang berdasarkan informasi pribadi dari konten-konten yang mereka buat [4].

Alat strategis yang dikembangkan berupa sistem yang dapat melakukan klasifikasi secara otomatis setiap postingan media sosial kandidat *programmer*. Hasil klasifikasi tersebut diharapkan mampu memprediksi pola kinerja dari masing-masing kandidat dengan hasil kinerja baik atau buruk. Indikator yang digunakan sebagai parameter kelas klasifikasi yaitu (1) update dunia teknologi, (2) *experiment*, (3) komunitas, (4) *share knowledge*, (5) portfolio, (6) *attitude*, (7) *mentoring*, (8) opini diskusi dan (9) promosi [5]. Metode klasifikasi yang dapat digunakan antara lain *Support Vector Machine* (SVM), *Random Forest* (RF) dan *Stochastic Gradient Descent* (SGD).

Penelitian yang dilakukan oleh Tu [4] menyatakan bahwa fitur jejaring sosial dapat digunakan untuk memprediksi kepribadian seseorang dengan akurasi hingga 78,6 % pada model kepribadian Big 5 *yaitu openness, conscientiousness, extraversion, agreeableness* dan *neuroticism*.

Penelitian yang dilakukan oleh Aliady [6] menunjukan bahwa kinerja dari metode RF lebih unggul daripada metode SVM dalam melakukan diagnosis penyakit kanker payudara dengan perbandingan nilai akurasi sebesar 94,5% untuk RF dan 93,1% untuk SVM.

Penelitian yang dilakukan oleh Maulina [7] tentang klasifikasi artikel hoax dengan *Linear* SVM dengan pembobotan *Term Frequency* dan *Inverse Document Frequen* dan menghasilkan akurasi klasifikasi dengan 108 artikel *hoax* dan 132 artikel tidak hoax adalah 95.3%.

Penelitian yang dilakukan oleh Oktanisa [8] tentang perbandingan metode klasifikasi data mining untuk direct bank menghasilkan SGD merupakan metode terbaik dengan tingkat akurasi sebesar 97,2 %.

Penelitian yang dilakukan oleh Ariadi [9] tentang klasifikasi berita berbahasa Indonesia dengan metode *Linear* SVM dan *Naïve Bayesian Classification* (NBC) menghasilkan *Linear* SVM lebih baik dibandingkan dengan NBC.

Penelitian ini bertujuan untuk mengetahui hasil akurasi terbaik dari ketiga metode yaitu SVM, RF dan SGD. Metode dengan akurasi terbaik akan dijadikan sebagai model klasifikasi otomatis dalam memprediksi kinerja *programmer* berdasarkan aktivitas media sosial.

2. Metode Penelitian

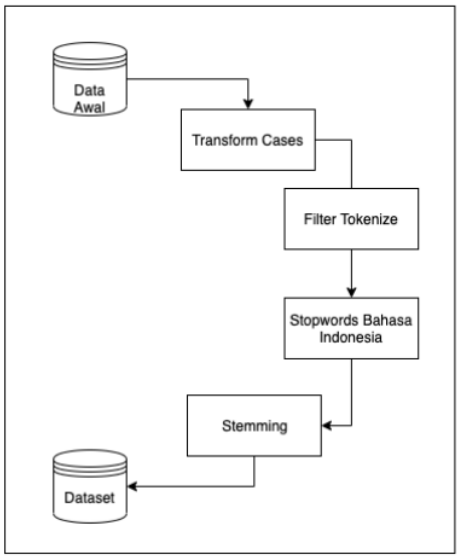
Beberapa tahap yang dilakukan untuk mencapai tujuan penelitian antara lain, (1) Tahap pengumpulan data, (2) *Text Preprocessing*, (3) Membagi data latih dan data uji, (4) Konfigurasi model klasifikasi, (4) Pembuatan model klasifikasi dengan SVM, RF dan SGD, (5) Uji validasi hasil, (6) Hasil perbandingan performa antar metode, dan (7) Kesimpulan.

2.1. Pengumpulan Data

Informasi yang baik diperlukan agar tujuan penelitian dapat tercapai. Metode pengumpulan data yang digunakan pada penelitian ini antara lain *survey*, observasi, studi pustaka dan studi literatur sejenis.

2.2. *Text Preprocessing*

Dataset yang diperoleh tidak bisa langsung diolah karena masih belum memiliki arti secara jelas (*missing value*), sehingga harus dilakukan pemilahan dan pembersihan data (*cleansing data*). *Text preprocessing* bertujuan untuk melakukan normalisasi data teks, pembersihan teks dari simbol-simbol, tanda baca dan karakter yang tidak diperlukan sehingga memiliki makna yang lebih jelas [10]. Tahapan dalam *text preprocessing* dapat dilihat pada Gambar 1.



Gambar 1. *Text Preprocessing*

Transform cases merupakan perubahan deretan kata dalam suatu teks menjadi bentuk kecil semua (*lower case*) atau dalam bentuk besar semua (*capital case*) [11]. *Tokenize* digunakan untuk menghilangkan berbagai karakter, simbol serta tanda baca yang dianggap tidak penting dengan cara melakukan penyaringan berdasarkan panjang suatu teks [11]. *Stop words* dilakukan cara menghilangkan kata-kata umum yang biasanya muncul dalam jumlah yang besar dan dianggap tidak memiliki makna serta tidak berpengaruh dalam proses klasifikasi kata. Contoh *stop words* dalam bahasa Indonesia seperti "yang", "di", "ke" [12]. *Stemming* ialah mencari akar kata (kata dasar) dari setiap kata yang telah diproses setelah *stop words* [13].

Hasil *text preprocessing* adalah *dataset* berupa file *microsoft excel* dengan format (.csv) dengan separator koma (,) sebagai pemisah antara postingan dengan kelasnya.

2.3. TF-IDF

TF-IDF dilakukan setelah melakukan tahap *text preprocessing*. TF-IDF ialah sebuah metode integrasi antara *term frequency* (TF), dan *inverse document frequency* (IDF). Metode TF-IDF berfungsi sebagai representasi nilai dari masing-masing dokumen dari kumpulan data latih sehingga terbentuk *vector* antara dokumen dengan kata. Rumus TF dihitung dengan menggunakan persamaan (1) dengan *term frequency* ke-i merupakan kemunculan *term* ke-*i* dalam dokumen ke-*j*. IDF yaitu logaritma dari rasio jumlah keseluruhan dokumen dalam korpus dengan jumlah dokumen yang mempunyai *term* yang dapat dilihat pada persamaan (2). Nilai TF-IDF dilakukan dengan cara mengalikan keduanya yang dapat dilihat pada persamaan (3) [14].

$$tf_i = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (1)$$

$$idf_i = log \frac{|D|}{|\{d: t_i \epsilon d\}|} \quad (2)$$

$$(tf - idf)_{ij} = tf_i(d_j) * idf_i \quad (3)$$

Keterangan:
tf_i = *term frequency*
f_i = *frequency*
d_j = *document*
t_i = *term*
idf_i = *inverse document frequency*
ϵd = himpunan *document*
D = jumlah semua *document*

2.4. Seleksi Fitur

Seleksi fitur dilakukan setelah melakukan pembobotan kata dengan metode TF-IDF. Nilai bobot dihasilkan dari setiap paragraf yang ada pada suatu dokumen. Kata-kata yang memiliki bobot tertinggi dari masing-masing paragraf kemudian diseleksi dan dipilih dan digabungkan mejadi sebuah ringkasan [14].

2.5. *Support Vector Machine*

SVM pada dasarnya berfungsi sebagai metode untuk menghasilkan garis pemisah (*hyperplane*) dari himpunan data ke dalam dua kelas secara *linier*. *Hyperplane* di sini ialah istilah yang dibuat secara umum untuk semua dimensi. Sebagai contoh, untuk himpunan data berdimensi satu, maka *hyperplane* dapat berwujud sebagai sebuah titik, jika himpunan dalam bentuk dua dimensi maka *hyperplane* berbentuk garis lurus dan jika berdimensi tiga maka berbentuk bidang datar [15]. Pencarian *hyperplane* paling optimum pada klasifikasi linier dapat dilakukan dengan persamaan (4) [16].

$$min \frac{1}{2}||\omega||^2 \quad (4)$$

Subject to:

$$y_i(wx_i + b) \geq 1, i = 1, \dots, \lambda \quad (5)$$

Keterangan:
x_i = Data input
w, b = Parameter yang akan dicari nilainya
y_i = Keluaran dari data xi
ω = Parameter model

Persamaan (5) digunakan unutk meminimalkan fungsi tujuan $\frac{1}{2}||\omega||^2$ atau memaksimalkan kuantitas $||\omega||^2$ dengan cara memperhatikan pembatas $y_i(wx_i + b) \geq 1$. Apabila output data $y_i = +1$, maka pembatas menjadi $(wx_i + b) \geq 1$ dan bila y_i = -1, pembatas menjadi $(wx_i + b) \geq -1$.

Metode SVM sering digunakan sebagai media klasifikasi yang bersifat otomatis seperti klasifikasi teks, citra, analisis medik ataupun prediksi. Klasifikasi ialah pengumpulan objek yang memiliki ciri khas yang sama ke dalam beberapa kelas. SVM dapat digunakan sebagai parameter yang digunakan dalam pembagian klasifikasi data [17].

Awalnya SVM dibuat hanya digunakan untuk mengklasifikasi dua buah kelas (*binary classifier*). Penelitian terus berlanjut hingga menjadikan SVM bersifat multi kelas (*multi classifier*) dengan kemampuan mengklasifikasi lebih dari dua kelas. Untuk klasifikasi data dalam *k* kelas maka diharuskan membuat $\frac{k(k-1)}{k}$ model SVM *Biner*. Sebagai contoh untuk membuat klasifikasi 4 kelas maka, harus membangun $\frac{4(4-1)}{4}$ = 6 buah SVM *Biner*. SVM *Biner* pertama dilatih dengan data latih dari kelas pertama dan kelas kedua untuk mengklasifikasikan data ke dalam C1 atau C2. SVM *Biner* kedua dilatih dengan data latih dari kelas C2 atau C3, dan seterusnya. Setiap kelas harus dibandingkan dengan tiga kelas laiinya. Cara *voting* dilakukan untuk menghasilkan kelas keputusan. Kelas yang paling sering menang adalah kelas keputusan [18].

2.6. *Random Forest*

Menurut [19] metode RF atau bisa disebut juga dengan *random esembles* yang berarti hutan acak yaitu kombinasi pohon keputusan sedemikian hingga setiap pohon bergantung terhadap nilai-nilai vector acak yang disampling secara independen dan dengan distribusi yang sama untuk semua pohon tersebut. Kelebihan dari metode ini adalah pada proses pemilihan fitur secara acak untuk memilah setiap simpul agar menghasilkan tingkat kesalahan yang relatif rendah.

Metode RF bekerja dengan cara menumbuhkan pohon sampai terbentuk hutan (*forest*). Kumpulan pohon yang terbentuk kemudian dilakukan analisa. Berdasarkan

himpunan data yang terdiri atas *n* data pantau dan p variabel penjelas, maka langkah-langkah dalam metode RF yaitu [6]:

1. Penarikan pada sampel acak berukuran *n* dengan pemulihan pada himpunan data (tahapan *bootstraping*).
2. Pohon dibangun hingga tinggi maksimum. Proses pemisahan dengan cara memilih secara acak jumlah variabel prediktor $(m) < d$ variabel penjelas, dan lakukan pemisahan terbaik.
3. Lakukan perulangan langkah 1 dan 2 dengan *k* kali sehingga terbentuk sebuah hutan yang terdiri atas *k* pohon acak.
4. Lakukan prediksi gabungan berdasarkan *k* buah pohon tersebut.

Nilai *m* merupakan salah satu variabel yang dapat diubah. Variabel ini digunakan untuk variabel penjelas yang digunakan sebagai kandidat pemisah dalam pembentukan pohon. Nilai *m* yang semakin besar akan menyebabkan korelasi semakin besar.

2.7. *Stochastic Gradient Descent*

Gradient Descent merupakan salah satu algoritma optimasi iteratif untuk menemukan titik yang meminimumkan suatu fungsi yang dapat diturunkan. Metode ini bekerja dengan memulai dari sebuah tebakan awal dan secara iteratif tebakan ini dapat diperbaiki berdasarkan suatu aturan yang melibatkan *gradient*/turunan pertama dari fungsi yang ingin diminimumkan. Persamaan (6) digunakan dalam kasus yang secara khusus mengatur langkah-langkah yang diambil untuk menurunkan fungsi yang ingin diminimumkan [16].

$$\omega_i + 1 = \omega_i - \eta \nabla_{\omega_i} L(\omega_i) \quad (6)$$

Keterangan:
$\omega_i + 1$ = Parameter model untuk prediksi
ω_i = Parameter model pada iterasi sebelumnya
η = Tingkat pembelajaran
L = Fungsi *loss cost*

Tahap pembelajaran dalam *Gradient Descent* standar mensyaratkan bahwa turunannya dihitung untuk semua sampel dalam *dataset* pelatihan disetiap iterasi. Konsekuensinya adalah ketika data latih terlalu besar dapat terjadi komputasi secara intensif. SGD merupakan salah satu varian dari *Gradient Descent* yang dilatih sebagai sampel pelatihan tunggal yang dipilih secara acak pada suatu waktu. SGD bersifat lebih skalabel dan cepat untuk dilatih dengan tanpa adanya batasan waktu dalam pelaksanaan dengan ukuran *dataset* pelatihan. Tujuan penggunaan SGD yaitu pelaksanaan klasifikasi yang lebih cepat dalam pembelajarannya. Fungsi *hinge loss* yang digunakan untuk melatih *classifier* dapat dilihat dengan persamaan (7) [16].

$$L(x_j, y_j) = \max(0,1 - y_j.(\omega x_j + b)) \quad (7)$$

Keterangan:
$\omega \; dan \; b$ = Parameter model untuk prediksi
x_j = Sampel input
y_j = Kelas target

Fungsi ini digunakan sebagai metrik klasifikasi yang menilai model linier yang diprediksi dengan SGD disetiap iterasi pada fase pembelajaran, dan memodifikai dua parameternya (ω, b) sesuai dengan persamaan (8). Dalam kasus klasifikasi SGD, ω sesuai dengan bobot yang ditetapkan untuk fitur hamburan balik dalam fungsi keputusan, dan b adalah intersepnya. Properti menarik dari fungsi *hinge loss* adalah ia akan menghukum baik sampel yang salah dalam klasifikasi namun diberikan kepercayaan yang rendah sebagai pembatas antar kelas.

Istilah regularisasi ditambahkan ke fungsi kerugian (*loss function*) dalam persamaan (7) untuk membantu model yang diprediksi dan mengeneralisasi ke data yang tidak berlabel. Idenya adalah untuk menghukum model kompleks yang cenderung *overfitting*, yang ditandai dengan nilai yang lebih besar untuk parameter ω_i. Persamaan umum untuk regularisasi terdapat pada rumus (8) dan (9).

$$L1 = \sum_{i=1}^{m} |\omega_i| \quad (8)$$

$$L2 = \sum_{i=1}^{m} \omega_i^2 \quad (9)$$

Keterangan:
m = variabel prediktor
ω = Parameter model untuk prediksi

Nilai optimal untuk hiperparameter yang terdapat pada Tabel 1, yang merupakan jumlah iterasi SGD, loss function, istilah reularisasi dan koefisien α, dapat ditentukan dengan *cross-validation* di mana *classifier* dilatih pada satu bagian dari dataset pelatihan dengan nilai-nilai berbeda dari hiperparameter ini, dan kemudian divalidasi pada sisa dari *dataset* yang sama. Himpunan nilai-nilai hiperparameter menghasilkan skor akurasi terbaik selama *cross-validation* dapat digunakan untu pelatihan selanjutnya pada seluruh *dataset* pembelajaran dengan dua *dataset* yang berbeda, tidak ada peningkatan daam hal akurasi yang diamati dengan *cross-validation*. Selanjutnya, demi waktu komputasi yang lebih cepat, nilai-nilai hiperparameter standar SGD disimpan dan dilaporkan pada Tabel 1 kecuali untuk jumlah iterasi yang meningkat menjadi 1000 iterasi.

Tabel 1. Nilai-Nilai Hiperparameter Dalam Pelatihan SGD

Nomor Iterasi	Loss Function	Aturan Regularisasi	Alpha
1000	Hinge Loss	12	0,0001

2.7. *Confusion Matrix*

Menurut [15] Evaluasi model klasifikasi memerlukan data uji yang tidak digunakan dalam pelatihan. Evaluasi dapat dilakukan dengan suatu ukuran tertentu yaitu (1) *Accuracy*, (2) *Recall* (3) *Precission* (3) dan (4) *F1-Score*. Penjelasan TP, TN, FP dan FN adalah sebagai berikut:

1. TP (*True Positive*) yaitu jumlah tuple positif dilabeli dengan benar oleh model klasifikasi.
2. TN (*True Negative*) yaitu jumlah tuple negatif yang dilabeli dengan benar oleh model klasifikasi.
3. FP (*False Positive*) yaitu jumlah tuple negatif yang salah dilabeli oleh model klasifikasi.
4. FN (*False Negative*) yaitu jumlah tuple postif yang dilabeli salah oleh model klasifikasi.

Ukuran evaluasi model klasifikasi, di mana TP adalah *true positives*, TN adalah *true negative*, FP adalah *false positives*, FN adalah *false negative*, P adalah sampel positif dan N adalah jumlah sampel negatif. *Accuracy* atau disebut dengan tingkat pengenalan dapat dihitung dengan Rumus (10).

$$\frac{TP + TN}{P + N} \quad (10)$$

Recall atau disebut dengan *true positive rate* atau *sensitivity* dapat dihitung dengan Rumus (11).

$$\frac{TP}{P} \quad (11)$$

Precission atau ukuran kepastian label data positif dapat dihitung dengan Rumus (12).

$$\frac{TP}{TP + FP} \quad (12)$$

F1-Score atau rata-rata harmonik dari precission dan recall dapat dihitung dengan Rumus (13).

$$\frac{2\ x\ precission\ x\ recall}{precission + recall} \quad (13)$$

2.8. *Cross Validation*

Validasi dilakukan dengan metode *Cross Validation*. *Cross Validation* adalah sebuah metode untuk mempredikisi keakuratan data pengujian [20]. Metode *Cross Validation* yang digunakan adalah *K- fold cross validation* yaitu, sebuah teknik yang dapat melakukan pengulangan data latih (*training)* dan data uji (*testing)* dengan sebanyak *k* pengulangan dan pembagian *1/k* dari dataset, dimana *1/k* tersebut akan digunakan sebagai data uji [21]. Sebagai analogi misalkan kita memiliki 500 data. Kita asusmsikan k = 5 maka, keseluruhan data akan dibagi menjadi lima lipatan sehingga masing-masing memiliki 100 data. Setiap lipatan yang memiliki 100 data terlebih dahulu ditentukan mana yang sebagai data latih dan uji. Perbandingan data latih dan uji adalah 75%: 25% sehingga, 75 data latih dan 25 data uji. Berdasarkan kelima lipatan maka, terdapat 4 lipatan yaitu 4 x 100 = 400 data *training* dan, sisanya yaitu 100 sebagai lipatan data uji.

3. Hasil dan Pembahasan

3.1. Dataset Postingan *Programmer*

Dataset diperoleh dengan melakukan observasi dan *survey* dari anggota group *facebook* PHP Indonesia dalam bentuk postingan kandidat *programmer* terpilih. Jumlah dataset yang berhasil dikumpulkan sebanyak 2178 postingan *facebook*. *Dataset* disimpan dalam format (.csv) menggunakan Microsoft Excel dengan separator koma (,) yaitu Posting dan Kategori. Posting merupakan postingan *facebook* kandidat, sedangkan Kategori ialah parameter yang merupakan indikator kinerja *programmer* dari aktivitas media sosial [5]. Peringkasan indikator *share knowledge, experiments* dan update dunia teknologi dijadikan satu dengan nama "*share knowledge & experiments*" karena isi postingan dari ketiga indikator tersebut masih saling terkait satu sama lain sehingga terdapat tujuh indikator yang digunakan yaitu *attitude*, komunitas, promosi, *portfolio*, *share knowledge & experiments*, *mentoring* dan opini diskusi. *Dataset* dapat diakses secara *public* pada alamat *https://www.kaggle.com/purwonopurwono/dataset-postingan-programmer-facebook*.

3.2. Performa *Support Vector Machine*

Dataset terlebih dahulu dibagi menjadi dua bagian dengan 70% data latih dan 30% data uji. Parameter yang digunakan dalam SVM yaitu *random_state* sebagai pembuat bilangan acak dalam proses klasifikasi dan *tol* sebagai nilai toleransi kapan kriteria akan berhenti [22]. Metode ini menggunakan nilai *random_state*=0 dan *tol*=1e-4. Hasil *precission, recall*, *f1-score* dan *accuracy* pada klasifikasi menggunakan algoritma SVM ditunjukan pada Tabel 2.

Tabel 2. Hasil Confusion Matrix SVM

Indikator	Precission	Recall	F1-Score	Accuracy
Attitude	0,90	0,80	0,85	0,8
Komunitas	0,82	0,89	0,85	0.88652482
Promosi	0,74	0,78	0,76	0.78181818
Portfolio	0,84	0,84	0,84	0.84210526
Share Knowledge & Experiments	0,85	0,64	0,73	0.63934426
Mentoring	0,00	0,00	0,00	0,0
Opini Diskusi	0,87	0,77	0,83	0.77419355

3.3. Performa *Random Forest*

Performa dari algoritma RF diteliti hasilnya setelah melakukan pengujian menggunakan algortima SVM. Pembagian data latih dan data uji sama dengan uji sebelumnya yaitu 70% berbanding 30%. Parameter yang digunakan dalam RF yaitu *n_jobs* yang merupakan jumlah pekerjaan yang akan dilakukan secara paralel dalam membuat pohon acak [22] dan

random_state. Metode ini menggunakan nilai *random_state*=0 dan *n_jobs*=2. Hasil *precission, recall, f1-score* dan *accuracy* pada klasifikasi menggunakan algoritma RF ditunjukan pada Tabel 3.

Tabel 3. Hasil Confusion Matrix RF

Indikator	Precission	Recall	F1-Score	Accuracy
Attitude	0,86	0,78	0,82	0.89090909
Komunitas	0,86	0,78	0,82	0.78368794
Promosi	0,69	0,75	0,72	0.75151515
Portfolio	0,76	0,68	0,72	0.68421053
Share Knowledge & Experiments	0,96	0,41	0,57	0.40983607
Mentoring	0,00	0,00	0,00	0.0
Opini Diskusi	0,42	0,74	0,53	0.74193548

3.4. Performa *Stochastic Gradient Descent*

Algoritma ketiga yang diuji performanya setelah SVM dan RF adalah SGD. Pembagian data latih dan uji juga dengan komposisi yang sama yaitu 70% berbanding 30%. Parameter yang digunakan dalam SGD yaitu *loss* sebagai penghitung kesalahan model dalam proses optimasi, *penalty* sebagai ukuran kesalahan dalam memprediksi data klasifikasi, *random_state* sebagai pembuat bilangan acak, *max_iter* sebagai pembatas data latih dan *tol* sebagai penanda kapan kriteria akan berhenti. Metode ini menggunakan nilai *loss = 'hinge', penalty='l2', alpha=1e-3, random_state=42, max_iter=5* dan *tol=none*. Hasil *precission, recall, f1-score* dan *accuracy* pada klasifikasi menggunakan algoritma SGD ditunjukan pada Tabel 4.

Tabel 4. Hasil Confusion Matrix SGD

Indikator	Precission	Recall	F1-Score	Accuracy
Attitude	0,94	0,80	0,86	0.8
Komunitas	0,82	0,90	0,86	0.89716312
Promosi	0,73	0,77	0,75	0.76969697
Portfolio	0,84	0,84	0,84	0.84210526
Share Knowledge & Experiments	0,85	0,56	0,67	0.55737705
Mentoring	0,00	0,00	0,00	0.0
Opini Diskusi	0,83	0,79	0,81	0.79032258

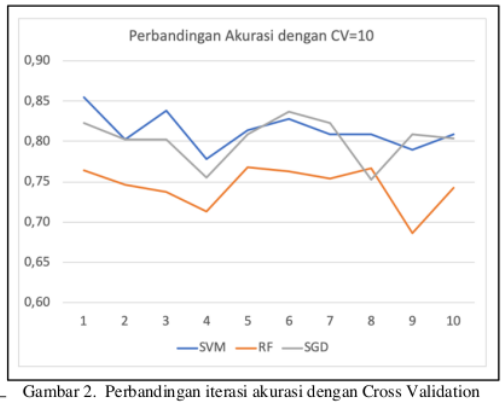
3.5. Hasil Validasi

Validasi hasil menggunakan metode k=10 *cross validation* dilakukan untuk melihat akurasi terbaik dari ketiga algoritma. Perbandingan rata-rata akurasi (*mean accuracy*) dari ketiga algoritma dapat dilihat secara detail pada Tabel 5.

Tabel 5. Perbandingan Akurasi Algoritma Dengan CV=10

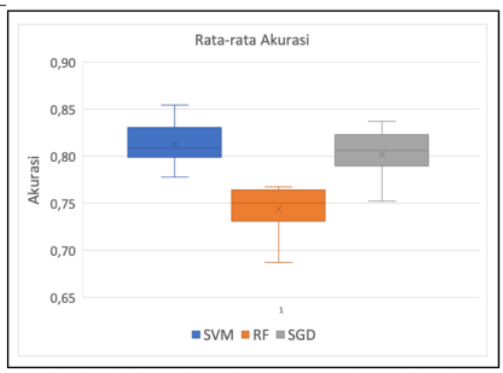
Pengujian Ke	SVM	RF	SGD
1	0,854545455	0,763636364	0,822727273
2	0,801843318	0,746543779	0,801843318
3	0,838709677	0,737327189	0,801843318
4	0,777777778	0,712962963	0,75462963
5	0,813953488	0,76744186	0,809302326
6	0,827906977	0,762790698	0,837209302
7	0,809302326	0,753488372	0,823255814
8	0,808411215	0,76635514	0,752336449
9	0,789719626	0,686915888	0,808411215
10	0,808411215	0,742990654	0,803738318

Berdasarkan Tabel 5 nilai akurasi tertinggi yang didapatkan oleh algoritma SVM yaitu sebesar 85,4% pada pengujian ke-1, namun memiliki nilai akurasi yang lebih rendah dibandingan dengan penelitian yang dilakukan oleh Aliady [6] yaitu 93,1% dan penelitian yang dilakukan oleh Maulina [7] dengan nilai 95,3%. Hasil implementasi algoritma RF menghasilkan akurasi tertinggi sebesar 76,3% pada pengujian ke-1 namun memiliki akurasi yang lebih rendah dari penelitian Aliadi [6] yaitu sebesar 94,5%. Hasil implementasi algoritma SGD mendapatkan nilai akurasi tertinggi sebesar 83,7% pada pengujian ke-6 dan belum bisa melampaui akurasi dari penelitian yang dilakukan oleh Oktanisa [8] yang bernilai 97,2 %. Perbandingan iterasi akurasi dari ketiga algoritma dapat dilihat lebih mudah dengan penggunaan grafik *line* pada Gambar 2.



Gambar 2. Perbandingan iterasi akurasi dengan Cross Validation

Perbandingan rata-rata akurasi dari ketiga algoritma tersebut jika disajikan dalam bentuk grafik *histogram* dapat dilihat pada Gambar 3 dimana terlihat metode SVM memiliki nilai akurasi tertinggi yaitu sebesar 85,4% diikuti oleh metode SGD dengan nilai akurasi tertinggi sebesar 83,7 % dan terakhir adalah metode RF dengan nilai akurasi tertinggi sebesar 76,3%.



Gambar 3. Perbandingan Mean Accuracy dengan CV=10

Perbandingan rata-rata akurasi dari ketiga algoritma juga dapat dilihat secara detail pada pada Tabel 4

dimana rata-rata tertinggi dimiliki oleh metode SVM dengan nilai 81,3% diikuti oleh SGD dengan nilai 80,1% dan RF dengan nilai 74,4%.

Tabel 6. Perbandingan Mean Accuracy Dengan CV=10

Model Name	Mean Accuracy	Standard Deviation
SVM	0,813058	0,022541
RF	0,744045	0,026145
SGD	0,801530	0,027743

Berdasarkan Tabel 6, ukuran rata-rata dari akurasi menggunakan k=10 dari *cross validation*, didapatkan hasil bahwa performa algoritma SVM lebih unggul daripada algoritma RF dan SGD dalam klasifikasi ini.

4. Kesimpulan

Berdasarkan hasil penelitian didapatkan bahwa perbandingan performa dari tiga metode *machine learning* yaitu SVM, RF dan SGD dalam klasifikasi kinerja *programmer* pada aktivitas media sosial didapatkan algoritma dengan performa terbaik yaitu SVM yang rata-rata tingkat akurasinya lebih tinggi daripada RF dan SGD. Hasil klasifikasi menunjukan persentase rata-rata akurasi dengan k=10 *cross validation* untuk algoritma SVM mencapai angka 81,3%, RF pada angka 74,4 %, dan SGD pada angka 80,1%.

Kontribusi penelitian ini adalah menghasilkan keputusan dalam penentuan model klasifikasi yaitu metode SVM. Metode SVM dapat digunakan pada dataset postingan media sosial yang berbeda. Penelitian selanjutnya ialah melakukan klasifikasi postingan berbentuk gambar atau video dari aktivitas media sosial kandidat *programmer*. Hasil akhir dari penelitian ini diharapkan menghasilkan sebuah sistem berbasis website atau mobile yang mampu melakukan klasifikasi otomatis konten media sosial dari setiap kandidat *programmer*. Sistem tersebut dapat dimanfaatkan oleh pihak rekrutmen sebagai sarana mendapatkan kandidat-kandidat terbaik untuk perusahaan berbasis teknologi informasi.

Daftar Rujukan

[1] M. A. Jaya, R. Ferdiana, and S. Fauziati, "Analisis Faktor keberhasilan Startup Digital di Yogyakarta," in *Prosiding ... TIF*, 2017, vol. 4, no. 1, pp. 167–173.

[2] M. D. K. Perdani, Widyawan, and P. I. Santoso, "Faktor-faktor yang mempengaruhi pertumbuhan startup di yogyakarta," in *Seminar Nasional Teknologi Informasi dan Komunikasi 2018* ..., vol. 2018, no. Sentika, pp. 23–24.

[3] T. Koch, C. Gerber, and J. J. De Klerk, "The impact of social media on recruitment: Are you Linkedin?," *SA J. Hum. ... pur. Manag.*, vol. 16, pp. 1–14, 2018.

[4] C. Tu, Z. Liu, H. Luan, and M. Sun, "PRISM: Profession identification in social media," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 6, 2017.

[5] Purwono, R. Umar, and I. Riadi, "Perancangan Indikator Kinerja Programmer pada Aktivitas Media Sosial," *J. ... ilkom UMB*, 2020.

[6] H. Aliady *et al.*, "Implementasi Support Vector Machine (...) Dan Random Forest pada Diagnosis Kanker Payudara," *Semin. Nas. Teknol. Inf. dan Komun. 2018 (SENTIKA 2018)*, ... 2018, no. Sentika, pp. 278–285, 2018.

[7] D. Maulina and R. Sagara, "Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency – Inverse Document ... equency," *Mantik Penusa*, vol. 2, no. 1, pp. 35–40, 2018.

[8] I. Oktanisa and A. A. Supianto, "Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank a Comparison of Classification Techniques in Data Mining for," *Teknol. Inf. ... n Ilmu Komput.*, vol. 5, no. 5, pp. 567–576, 2018.

[9] D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *J. SAINS DAN SENI ITS Vol. 4, No.2*, vol. 4, no. 2, pp. 248–253, 2015.

[10] A. Handayanto, K. Latifa, N. D. Saputro, and R. R. Waliyansyah, "Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi (Analysis and Application of Algorithm Support Vector Machine ... SVM) in Data Mining to Support Promotional Strategies)," *JUITA J. Inform.*, vol. 7, ... November, pp. 71–79, 2019.

[11] A. T. J. H, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *Inform. UPGRIS*, vol. 1, pp. 1–9, 2015.

[12] A. Alajmi and E. mostafa Saad, "Toward an ARABIC Stop-Words List Generation Toward an ARABIC Stop-Words List Generation," no. January 2012, 2018.

[13] J. Shodiq and L. A. Muharom, "Kategorisasi Dokumen Text Menggunakan Metode K-Nearest Neighbor pada Dokumen Tugas Akhir Universitas Muhammadiyah Jember," Universitas ... hammadiyah Jember, 2017.

[14] N. K. Widyasanti, I. K. G. D. Putra, and N. K. D. Rusjayanthi, "Seleksi Fitur Bobot Kata dengan Metode TF-IDF untuk Ringkasan Bahasa Indonesia," *Merpati*, vol. 6, no. 2, pp. 119–... 6, 2018.

[15] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*. ... dung: Informatika, 2018.

[16] A. S. Ritonga and E. S. Purwaningsih, "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas ... ngelasan Smaw (Shield Metal Arc Welding)," *Ilm. Edutic*, vol. 5, no. 1, pp. 17–25, 2018.

[17] I. Riadi, R. Umar, and F. D. Aini, "Analisis Perbandingan Detection Traffic Anomaly Dengan Metode Naive Bayes Dan Support Vector Machine (Svm)," *Ilk. J. Ilm.*, vol. 11, no. 1, p. ... , 2019.

[18] N. I. Widiastuti, E. Rainarli, and K. E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen," *J. ... otel*, vol. 9, no. 4, p. 416, 2017.

[19] Suyanto, *Machine Learning Tingkat Dasar dan Lanjut*. ... dung: Informatika, 2018.

[20] T. H. Apandi, C. A. Sugianto, and C. R. Service, "Algoritma Naive Bayes untuk Prediksi Kepuasan Pelayanan Perekaman e-KTP (Naive Bayes Algorithm for Satisfaction Prediction of e-ID)," *JUITA J. Inform.*, vol. 7, no. November, pp. 125–128, ...

[21] S. Asiyah and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine Dan K-... arest Neighbor," *J. Sains dan Seni ITS*, vol. 5, no. 2, 2016.

[22] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn: Machine Learning in Python Fabian," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 2825–2830, 2011.

HASIL CEK_17 Perbandingan

ORIGINALITY REPORT

17%	16%	11%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Bina Sarana Informatika Student Paper	3%
2	widuri.raharja.info Internet Source	2%
3	jurnal.polibatam.ac.id Internet Source	1%
4	elibrary.unikom.ac.id Internet Source	1%
5	sisfotenika.stmikpontianak.ac.id Internet Source	1%
6	jurnal.darmajaya.ac.id Internet Source	1%
7	journal.universitasbumigora.ac.id Internet Source	<1%
8	repository.its.ac.id Internet Source	<1%
9	www.ojs.stmikpringsewu.ac.id Internet Source	<1%

10	Cunchao Tu, Xiangkai Zeng, Hao Wang, Zhengyan Zhang, Zhiyuan Liu, Maosong Sun, Bo Zhang, Leyu Lin. "A Unified Framework for Community Detection and Network Representation Learning", IEEE Transactions on Knowledge and Data Engineering, 2019 Publication	<1 %
11	Submitted to Udayana University Student Paper	<1 %
12	jurnal.mdp.ac.id Internet Source	<1 %
13	eprints.uad.ac.id Internet Source	<1 %
14	jurnal.umsu.ac.id Internet Source	<1 %
15	repository.usd.ac.id Internet Source	<1 %
16	openlibrarypublications.telkomuniversity.ac.id Internet Source	<1 %
17	djournals.com Internet Source	<1 %
18	jurnal.untan.ac.id Internet Source	<1 %
19	ru.scribd.com Internet Source	<1 %

20	stmik-budidarma.ac.id Internet Source	<1 %
21	Submitted to Universitas Islam Majapahit Student Paper	<1 %
22	Submitted to Universitas Siswa Bangsa Internasional Student Paper	<1 %
23	Namita Ruparel, Amandeep Dhir, Anushree Tandon, Puneet Kaur, Jamid Ul Islam. "The influence of online professional social media in human resource management: A systematic literature review", Technology in Society, 2020 Publication	<1 %
24	etd.repository.ugm.ac.id Internet Source	<1 %
25	e-jurnal.pelitanusantara.ac.id Internet Source	<1 %
26	csrid.potensi-utama.ac.id Internet Source	<1 %
27	Annisa Elfina Augustia, Resi Taufan, Yuris Alkhalifi, Windu Gata. "Analisis Sentimen Omnibus Law Pada Twitter Dengan Algoritma Klasifikasi Berbasis Particle Swarm Optimization", Paradigma - Jurnal Komputer dan Informatika, 2021	<1 %

	Publication	
28	dspace.uii.ac.id Internet Source	<1 %
29	ejournal.stmik-sumedang.ac.id Internet Source	<1 %
30	garuda.ristekbrin.go.id Internet Source	<1 %
31	repository.uin-suska.ac.id Internet Source	<1 %
32	repository.yudharta.ac.id Internet Source	<1 %
33	text-id.123dok.com Internet Source	<1 %
34	Rahmiati Rahmiati, Dedy Irfan, Agustin Agustin, Siska Hediyati. "APLIKASI PENGUKUR TINGKAT SENTIMEN PELANGGAN BERDASARKAN KOMPLAIN PELANGGAN PLN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR", INOVTEK Polbeng - Seri Informatika, 2020 Publication	<1 %
35	online-journal.unja.ac.id Internet Source	<1 %
36	Submitted to Sriwijaya University Student Paper	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		