Analisis Akurasi Penerapan Algoritma Support Vector Machine Menggunakan Kernel Radial Basis Function pada Penentuan Kelayakan Kredit (Studi Kasus German Kredit Data)

Danarko Pakuan Putra[1] dan Bheta Agus Wardijono[2]

[1]Magister Manajemen Sistem Informasi Universitas Gunadarma
[1]Jl Margonda Raya 100 Depok
[2]Sistem Komputer STMIK Jakarta STI&K
[2]Jl BRI Radio Dalam Jakarta Selatan
E-mail : artileri45@gmail.com, bheta@jak-stik.ac.id

Abstrak

Machine Learning merupakan disiplin ilmu kecerdasan buatan dengan teknik statistika untuk menghasilkan suatu model dari sekumpulan data. Terdapat banyak algoritma *machine learning* yang biasa digunakan, salah satunya *support vector machine* (SVM). SVM salah satu metode yang dapat melakukan pengklasifikasi data dengan baik, karena proses yang akan dilakukan bersifat non linear maka akan menggunakan kernel *radial basis function* (RBF). Fungsi RBF dapat diterapkan dalam mengklasifikasi permohonan kredit. Penelitian ini menggunakan German Credit Dataset. German Credit Dataset dengan 1000 data memiliki 21 variabel terdiri 20 variabel input dan 1 variabel target dengan kelas tidak seimbang. Hasil dari pengujian ditampilkan dalam bentuk, confusion matrix yang akan digunakan untuk perhitungan akurasi dan area *under curve* untuk perhitungan performa. Dari percobaan yang dilakukan diperoleh nilai akurasi 0.743 (74%) dan nilai performa 0.7689 (77%) termasuk kategori *Fair Model*.

Kata Kunci : *Support Vector Machine, Radial Basis Function*, German Credit Dataset

Pendahuluan

Perkembangan teknologi di dunia saat ini mengarah pada makin meluasnya pemanfaatan kecerdasan buatan atau *Artificial Intelegent* (AI). Salah satu algoritma yang digunakan pada AI adalah *Support Vector Machine* (SVM). SVM digunakan sebagai teknik untuk menemukan *hyperplane* yang memisahkan dua set data dari dua kelas yang berbeda dengan margin terbesar [1]. *Hyperplane* adalah garis batas yang memisahkan data antar-kelas. Margin adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas [2]. Kelebihan SVM adalah tidak mengalami overfitting karena training perlu dilakukan sekali saja dan mendapatkan solusi optimal [3]. Algoritma SVM memiliki beberapa fungsi kernel salah satunya kernel *Radial Basis Function* (RBF). RBF sangat cocok di gunakan untuk memecahkan masalah data yang non linear seperti German Credit Dataset. *Dataset* yang digunakan dalam penelitian ini diperoleh dari UCI *Repository of Machine Learning Datasets*. Tujuan dari penelitian ini yaitu untuk mengetahui akurasi dari algoritma *Support Vector Machine* dengan kernel *Radial Basis Function*.

Penelitian Terdahulu

Penelitian terdahulu merupakan tolak ukur peneliti untuk mencari perbandingan dan selanjutnya untuk menemukan inspirasi baru untuk penelitian selanjutnya di samping itu kajian

DOI : http://dx.doi.org/10.32409/jikstik.19.2.2786

terdahulu membantu penelitian dalam memposisikan penelitian serta menunjukkan orsinalitas dari penelitian.

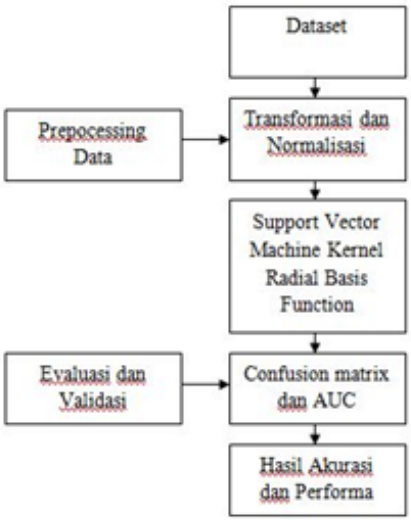
Penelitian pertama oleh Pusphita judul "Penerapan Metode Klasifikasi *Support Vector Machine* Pada Data Akreditasi Sekolah Dasar Di Kabupaten Magelang". Penelitian ini menyimpulkan dengan *Support Vector Machine Kernel Radial Basis Function* nilai akurasi data training 100% dan nilai akurasi data testing sebesar 93% [4]. Penelitian kedua oleh Rinawati dengan judul "Penentuan Penilaian Kredit Menggunakan Metode *Naive Bayes* Berbasis *Particle Swarm Optimization*". Penelitian ini menyimpulkan dengan menggunakan *Naïve Bayes* nilai akurasi dan performa adalah 72.40% dan 0.765 sedangkan menggunakan *Naïve Bayes* berbasis *Particle Swarm Optimization* nilai akurasi dan performa adalah 75.90% dan 0.773 [5]. Penelitian ketiga oleh Siti Harlina dengan judul "Analisa Data Mining Pada Penentuan Kelayakan Kredit Menggunakan Algoritma K-NN Berbasis *Forward Selection*". Penelitian ini menyimpulkan dengan menggunakan K-NN nilai akurasi yang didapat 61.90%, selanjutnya dengan menggunakan K-NN berbasis *Forward Selection* nilai akurasi yang didapat 73.60% [6]. Penelitian Keempat oleh Imelda dengan judul "Penerapan Metode *Support Vector Machine* (SVM) Menggunakan *Kernel Radial Basis Function* (RBF) Pada Klasifikasi Tweet". Penelitian ini menyimpulkan nilai akurasi 97,54%, sedangkan untuk data yang dilakukan feature nilai akurasi 99,12% [7]. Penelitian Kelima oleh Achmad Yani dengan judul "Analisa Kelayakan Kredit Menggunakan *Artificial Neural Network* dan *Backpropogation*". Penelitian ini menyimpulkan nilai akurasi 0.7133 (71%) dan nilai performa 0.72 (72%) [8].

Metode Penelitian

Metode penelitian diuraikan dalam alur penelitian untuk memberikan gambaran yang jelas, teratur dan sistematis. Penelitian ini dilakukan melalui sejumlah tahapan. Alur penelitian terlihat pada Gambar 1.

Dataset

Pada penelitian ini menggunakan *dataset* yang berasal dari *machine learning repository* yaitu german credit data yang disediakan oleh Professor Dr. Hans Hofmann Institut fuer Statistik und "Okonometrie University" at Hamburg [9]. Dataset ini memiliki 20 variabel input dan 1 variabel target dan berjumlah 1000 data kredit yang termasuk dalam jenis data kelas tidak seimbang, dimana 700 (70%) adalah data kredit yang baik dan 300 (30%) data kredit yang kurang baik. Data kredit German merupakan data publik karena data tersebut dipublikasikan dan data tersebut dapat diakses oleh siapa saja.



Gambar 1: Alur Penelitian

Preprocessing Data

Preprocessing data merupakan tahapan awal dan penting dalam melakukan proses klasifikasi data teks. Tujuan dilakukannya *text preprocessing* yaitu untuk menghilangkan *noise*, menyeragamkan bentuk kata dan mengurangi volume kata [10]. Peneliti menggunakan bahasa pemrograman dan perangkat lunak untuk analisis statistika dan grafik, yaitu program Ross dan Robert (R). Program R digunakan sebagai bantuan untuk metode transformation dan model normalisasi. Transformasi mengubah tipe atribut int menjadi numeric pada variable input dan tipe int menjadi *factor* pada *variable* target. Untuk model normalisai menggunakan metode min max.

Support Vector Machine Kernel Radial Basis Function

Support Vector Machine (SVM) merupakan sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik [11]. *Radial Basis Function* (RBF) merupakan fungsi *kernel* yang biasa digunakan dalam analisis ketika data tidak terpisah secara linear. RBF memiliki dua parameter yaitu *Gamma* dan *Cost*. Parameter Cost atau biasa disebut sebagai C merupakan parameter yang bekerja sebagai pengoptimalan SVM untuk menghindari misklasifikasi di setiap sampel dalam training *dataset*. Parameter *Gamma* menentukan seberapa jauh pengaruh dari satu sampel training *dataset* dengan nilai rendah berarti jauh, dan nilai tinggi berarti dekat [12]. Berikut fungsi RBF ditunjukan pada persamaan 1.

$$K(x_i, x) = exp(-y \parallel x_i - x \parallel^2, y > 0 \qquad (1)$$

Evaluasi dan Validasi

Validasi model dapat diukur dengan teknik *confusion matrix*. *Confusion matrix* merupakan visualisasi kinerja dari algoritma klasifikasi menggunakan data dalam matriks (Tabel 1). Hal tersebut membandingkan klasifikasi prediksi terhadap klasifikasi aktual dalam bentuk *False Positive* (FP) yang merupakan jumlah data positif namun terklasifikasi salah oleh sistem, *True Positive* (TP) jumlah data positif yang terklasifikasi dengan benar oleh sistem, *False Negative* (FN) yaitu jumlah data negatif namun terklasifikasi salah oleh sistem, dan *True Negative* (TN) adanya jumlah data negatif yang terklasifikasi dengan oleh sistem [13]. Rumus dari *confusion matrix* digunakan untuk menghitung *accuracy*, seperti ditunjukan pada persamaan 2.

Tabel 1: *Confussion Matrix*

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive count (TP)	False Negatives count (FP)
	Negative	False Positive count (FN)	*True Negatives count (TN)*

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad (2)$$

Untuk menghitung performa maka digunakan *Area Under Curve* (AUC) rumus AUC lihat persamaan 3.

$$AUC = \frac{1}{2} \sum_{i=1}^{n} (X_{i+1} - X_i)(Y_{i+1} - Y_i) \qquad (3)$$

AUC akan menghitung nilai yang berada dibawah grafik *Receiver Operating Characteristic* (ROC). Nilai ROC yang mendekati satu maka akan semakin baik seperti diperlihatkan pada Table 2.

Tabel 2: *Nilai Area Under Curve*

AUC	Interpretation
1.0 (100%)	*Perfect Model*
0.9 – 0.99 (90 - 99%)	*Excellent Model*
0.8 – 0.89 (80 – 89%)	*Very Good Model*
0.7 – 0.79 (70 – 79%)	*Fair Model*
0.51 – 0.69 (51 – 69%)	*Poor Model*
< 0.5 (50%)	*Worthless Model*

Analisis dan Pembahasan

Penelitian ini menggunakan program R. R adalah salah satu sistem perangkat lunak yang paling popular dan banyak digunakan untuk statistik, *data mining*, dan *machine learning* [14]. Pengujian model dilakukan menggunakan German Credit Dataset, untuk pelatihan dilakukan sebanyak 700 dan pengujian dilakukan sebanyak 300. Tahap awal dilakukan preprocessing data dengan cara transformasi dan normalisasi.

Hasil transformasi dan normalisasi terlihat pada Gambar 2 dan Tabel 3.

Tabel 3: Hasil Normalisasi

No	V1	V2	V3	V4	V5	V6	V7	V8	...	V21
1	A12	42+	03.A32.A33	A43	5500+	A61	A73	2	...	Bad
2	A14	06-12	04.A34	A46	1400-2500	A61	A74	2	...	Good
3	A11	36-42	03.A32.A33	A42	5500+	A61	A74	2	...	Good
...	...	...	...	...	...	...	...	...	...	...
1000	A11	12-24	03.A32.A33	A40	4500-5500	A61	A73	3	...	Bad

```
> str(cdata)
'data.frame':   1000 obs. of  21 variables:
 $ chk_ac_status_1                : Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4
 2 ...
 $ duration_month_2               : num  6 48 12 42 24 36 24 36 12 30 ...
 $ credit_history_3               : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3
 5 ...
 $ purpose_4                      : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2
 5 1 ...
 $ credit_amount_5                : num  1169 5951 2096 7882 4870 ...
 $ savings_ac_bond_6              : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4
 1 ...
 $ p_employment_since_7           : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4
 1 ...
 $ instalment_pct_8               : num  4 2 2 2 3 2 3 2 2 4 ...
 $ personal_status_9              : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1
 4 ...
 $ other_debtors_or_grantors_10   : Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1
 ...
 $ present_residence_since_11     : num  4 2 3 4 4 4 4 2 4 2 ...
 $ property_type_12               : Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3
 ...
 $ age_in_yrs_13                  : num  67 22 49 45 53 35 53 35 61 28 ...
 $ other_instalment_type_14       : Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3
 ...
 $ housing_type_15                : Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2
 ...
 $ number_cards_this_bank_16      : num  2 1 1 1 2 1 1 1 1 2 ...
 $ job_17                         : Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4
 ...
 $ no_people_liable_for_mntnance_18: num  1 1 2 2 2 2 1 1 1 1 ...
 $ telephone_19                   : Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
 $ foreign_worker_20              : Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
 $ good_bad_21                    : Factor w/ 2 levels "Bad","Good": 2 1 2 2 1 2 2 2 2 1 ...
>
```

Gambar 2: Hasil Transformasi

Data diatas adalah hasil dari transformasi yang mengubah data dengan tipe *int* menjadi *numeric* pada *variable input* dan tipe *int* menjadi *factor* pada variable target.

Tabel 2 adalah data hasil normalisasi data dengan menggunakan metode min max secara otomatisasi dari sistem R. Selanjutnya dilakukan pelatihan dengan 700 data dan pengujian 300 data.

Pelatihan telah berhasil dibuat dengan 21 variabel terdiri 20 Input dan 1 output dengan 700 dataset di lakukan secara acak. Hasil pelatihan terlihat pada tabel 4, berserta jumlah *Good* dan *Bad*.

Hasil pelatihan ini akan menjadi acuan untuk pengujian dengan menggunakan 300 data yang terbentuk pada confusion matrix dengan nilai terlihat pada Tabel 5.

Tabel 4: Data Hasil Pelatihan

No	V1	V2	V3	V4	V5	V6	V7	V8	...	V21
1	A12	00-06	04.A34	A43	0-1400	A65	A75	1	...	Good
2	A14	06-12	04.A34	A46	1400-2500	A61	A74	3	...	Good
3	A11	36-42	03.A32.A33	A42	5500+	A61	A74	4	...	Good
...	...	...	...	...	...	...	...	...	...	...
700	A14	00-06	02.A31	A40	0-1400	A65	A73	56	...	Bad

	Count	Percentage
Bad	210	30
Good	490	70

Tabel 5: *Confusion Matrix*

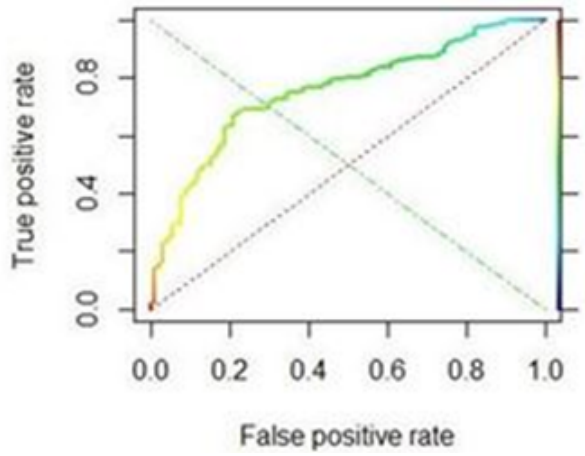
	Good	Bad
Good	43	30
Bad	47	180

Accuracy yang diperoleh adalah:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$= \frac{43 + 180}{43 + 180 + 30 + 47} = 0,743$$

Setelah mendapatkan nilai akurasi dari proses perhitungan confusion matrix, selanjutnya dicari perhitungan untuk mendapatkan performa yang divisualisasikan dengan kurva ROC. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.



Gambar 3: Kurva Receiver Operating Characteristic

Dari Gambar 3 terdapat grafik ROC dengan nilai AUC (Area Under Curve) sebesar 0.773 dimana diagnosa hasilnya *fair classification*.

Penutup

Pengujian model telah berhasil dilakukan menggunakan German Credit dataset yang diambil dari UCI *repository of machine learning*. Pengolahan data dilakukan dengan penerapan algoritma *Support Vector Machine kernel Radial Basis Function*, mendapatkan nilai akurasi berdasarkan perhitungan confusion matrix sebesar 0.743 (74%) dan untuk nilai performa metode mendapatkan nilai 0.773 (77%). Nilai performa terdapat pada range 70% sampai dengan 79% sehingga termasuk kedalam *Fair Model*, maka pengambilan keputusan bisa dilakukan walaupun tingkat akurasi masih kurang tinggi. Bagi peneliti selanjutnya bisa mengembangkan hasil penelitian ini dengan menggunakan metode yang lain dan komposisi data pelatihan yang berbeda.

Daftar Pustaka

[1] Y. Duan, J.S. Edwards & Y.K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data: evolution, challenges and research agenda", International Journal of Information Management, 48, 63-71, 2019.

[2] Sri Kusumadewi, "Artificial Intelligence (Teknik dan Aplikasinya)", Yogyakarta : Graha Ilmu, 2003.

[3] R. Abirami and M. S. Vijaya, "An Incremental Learning Approach for Stock Price Prediction Using Support Vector Regression", International Journal of Research and Reviews in Artificial intelligence (IJRRAI) Vol. 1, No. 4, 81-85, 2011

[4] Pusphita Anna Octaviani, Yuciana Wilandari, Dwi Ispriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang", Jurnal Gaussian, Vol.3, No.4, 811-820, 2014.

[5] Rinawati, "Penentuan Penilaian Kredit Menggunakan Metode Naïve Bayes Berbasis Particle Swarm Optimization", Jurnal Sains Komputer & Informatika, 1(1), 48-58, 2017.

[6] Sitti Harlina, "Data Mining Pada Penentuan Kelayakan Kredit Menggunakan Algoritma K-NN Berbasis Forward Selection", CCIT (Creative Communication and Innovative Technology) Journal, 11(2), 236-244, 2018.

[7] Imelda Amuis, Muhammad Affandes, "Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet", Jurnal Sains Teknologi dan Industri, 12(2), 189-197, 2015.

[8] Achmad Yani, Ega Hegarini, "Analisa Kelayakan Kredit Menggunakan Artificial Neural Network dan Backpropogation (Status Kasus German Credit Data)", Jurnal Ilmiah Komputasi, 18(4), 385-389, 2020.

[9] H. Hofmann,"Statlog (German Credit Dataset)", diakses daring pada https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german diakses tanggal 22 Oktober 2019 pukul 20.08 WIB.

[10] Han Jiawei., Micheline Kamber & Jian Pei, "Data Mining Concepts and Techniques", Elsevier Inc, 2012.

[11] Raudlatul Munawarah, Oni Soesanto, M. Reza Faisal, "Penerapan Metode Support Vector Machine Pada Diagnosa Hepatitis". Jurnal Ilmiah KLIK - Kumpulan Jurnal Ilmu Komputer, 3(1), 103-113, 2016.

[12] Hyeran Byun and Seong-Whan Lee, "A Survey on Pattern Recognition Applications of Support Vector Machines", International Journal of Pattern Recognition and Artificial Intelligence, Vol.17, 459-486, 2003.

[13] David L. Olson, and Dursun, Delen, "Advance Data Mining Techniques", German : Springer, 2008.

[14] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio and Z. M. Jones, "mlr : Machine Learning in R", Journal of Machine Learning Research, 17, 1-5, 2016.