

## Penerapan Model CRISP-DM pada Prediksi Nasabah Kredit yang Berisiko Menggunakan Algoritma *Support Vector Machine*

Tutut Wurijanto <sup>1)</sup>, Henry Bambang Setiawan <sup>2)</sup>, A.B. Tjandrarini <sup>3)</sup>

<sup>1)</sup>Program Studi S1 Sistem Informasi, Universitas Dinamika, email: [tutut@dinamika.ac.id](mailto:tutut@dinamika.ac.id)

<sup>2)</sup>Program Studi S1 Sistem Informasi, Universitas Dinamika, email: [henry@dinamika.ac.id](mailto:henry@dinamika.ac.id)

<sup>3)</sup>Program Studi D3 Sistem Informasi, Universitas Dinamika, email: [asteria@dinamika.ac.id](mailto:asteria@dinamika.ac.id)

### Abstrak

Prediksi klasifikasi nasabah kredit diperlukan untuk menentukan nasabah yang berisiko. Hal ini diperlukan agar pemberi hutang yakni bank tidak mengalami kredit macet. Kondisi saat ini hampir 46% nasabah mengalami kesulitan membayar hutang. Dengan adanya kasus tersebut maka penelitian ini bertujuan agar bank dapat memilah nasabah. Penelitian terkait dengan klasifikasi banyak dilakukan dengan bantuan data dari kaggle.com dan menggunakan atribut yang disesuaikan dengan kondisi tempat peneliti, yaitu: pendapatan, usia, pengalaman kerja, status pernikahan, kepemilikan rumah, kepemilikan mobil, lama bekerja, dan lama tinggal di rumah saat ini sebagai keputusan peminjaman. Parameter tersebut diproses dengan algoritma *Support Vector Machine* (SVM) dengan tujuan untuk klasifikasi nasabah yang berisiko dan tidak berisiko. Penelitian ini menggunakan sebanyak 100 data dengan pembagian 70% sebagai data latih dan 30% sebagai data tes. Selain itu, penelitian ini juga membandingkan SVM dengan Algoritma *Naive Bayes* dengan pembagian data yang sama. Penelitian menghasilkan nilai akurasi untuk SVM sebesar 80% dan *Naive Bayes* 73.33%.

**Kata Kunci:** Model CRISP-DM, Prediksi nasabah kredit, Algoritma *Support Vector Machine*

### PENDAHULUAN

Bank merupakan salah satu sumber modal yang dipakai oleh pengusaha dan pegiat Usaha Mikro Kecil dan Menengah (UMKM) untuk mengembangkan usaha. Hampir 46% pelaku usaha mengalami kesulitan membayar tagihan atau hutang (Azizah, 2022). Hal ini yang menyebabkan bank harus selektif untuk menyetujui permintaan hutang dari pelaku usaha untuk dipakai menambah modal bisnisnya. Apabila kejadian hutang macet terus berulang akan mengakibatkan bank mengalami banyak kesulitan dalam perputaran uangnya. Hal tersebut dapat dicegah apabila dilakukan analisis terjadinya hutang macet, dengan cara menggali data (*mining*) histori pelaku usaha yang akan berhutang.

Penggalan data nasabah atau peminjam untuk modal/kredit masih menarik untuk diteliti, apalagi dengan kondisi inflasi perekonomian saat ini yang

sulit diprediksi. Hal ini tentu menjadi perhatian penting bagi sektor pembiayaan atau lembaga perkreditan, salah satunya lembaga perbankan dalam mengidentifikasi dan memprediksi data nasabah yang berisiko sebagai pijakan dalam pengambilan keputusan oleh pihak manajemen.

Terdapat beberapa algoritma atau metode klasifikasi *data mining* yang dapat digunakan sebagai strategi pemasaran dan promosi, di antaranya *Support Vector Machines* (SVM), *Naive Bayes* (NB), dan *Decision Tree* (DT). Metode SVM merupakan salah satu metode pembelajaran mesin (*supervised learning*) yang dapat mengklasifikasikan data histori dengan mencari bidang pemisah atau *hyperplane* terbaik yang memisahkan data berdimensi tinggi secara sempurna ke dalam kelas-kelas (Zainuddin dan Selamat, 2014).

Prinsip kerja dari metode SVM ini adalah mencari ruang pemisah yang paling

optimal dari suatu *dataset* dalam kelas yang berbeda. *Hyperplane* dapat ditemukan dengan memaksimalkan *margin* atau jarak antara titik kelas terdekat (*support vector*) dan *hyperplane*. Data sampel seringkali tidak dipisahkan secara linier, namun SVM memperkenalkan gagasan untuk meningkatkan dimensi data. Pada umumnya penggunaan dimensi ruang yang lebih tinggi akan menyebabkan masalah mesin dan *overfitting*. Masalah tersebut dapat diselesaikan dengan penggunaan *dot-product* dalam ruang (Boswell, 2003).

Beberapa penelitian terkait *data mining* untuk tujuan prediksi telah dilakukan menggunakan metode SVM. Prediksi menggunakan metode SVM dilakukan oleh Iskandar dan Nataliani (2021), menghasilkan tingkat akurasi sebesar 96,43%. Penelitian sejenis tentang perbandingan dengan metode lain dilakukan oleh Pertiwi (2019) menghasilkan tingkat akurasi sebesar 89.03% dan oleh Iskandar dan Nataliani (2021) menghasilkan tingkat akurasi sebesar 93%.

Berdasarkan latar belakang tersebut masih terdapat peluang untuk melakukan penelitian sejenis dengan obyek yang berbeda. Penelitian ini diusulkan untuk mengklasifikasi nasabah bank yang berisiko dan tidak berisiko untuk diberikan pinjaman modal menggunakan algoritma SVM dan metode *Naive Bayes* yang mengacu pada sebuah standar proses *data mining* CRISP-DM. Kedua metode tersebut digunakan untuk dibandingkan kinerja atau tingkat akurasi. *Dataset* penelitian ini merupakan data publik berskala besar yang diproses sesuai tahapan *data mining* untuk menemukan pola sebagai dasar klasifikasi nasabah kredit bank yang berisiko atau tidak berisiko.

## METODE

### *Cross-Industry Standard Process for*

## *Data Mining*

Tahapan penelitian ini dilakukan dengan cara mengadopsi sebuah standar proses *data mining* yang disebut *Cross-Industry Standard Process for Data Mining* atau CRISP-DM. Standar proses tersebut terdiri atas lima fase yaitu fase pemahaman bisnis, fase pemahaman data, fase pengolahan data, fase pemodelan, dan fase evaluasi & validasi.



Gambar 1. Standar Proses Model CRISP-DM (Sumber: <https://www.datascience-pm.com/crisp-dm-2/>)

### 1. Fase Pemahaman Bisnis

*Dataset* penelitian merupakan data publik yang diunduh dari kaggle.com tentang prediksi nasabah berisiko.

### 2. Fase Pemahaman Data

*Dataset* tersebut terdiri atas 11 atribut prediktor dan 1 label dengan penjelasan yang terdapat dalam Tabel 1.

Tabel 1. Keterangan Atribut Prediktor

Variabel	Keterangan
<i>Income</i>	Pendapatan
<i>Age</i>	Usia
<i>Experience</i>	Pengalaman Kerja Keseluruhan
<i>Married/Single</i>	Status Pernikahan
<i>House_Ownership</i>	Status Kepemilikan Rumah
<i>Car_Ownership</i>	Status Kepemilikan Mobil
<i>Profession</i>	Profesi Saat Ini
<i>CITY</i>	Kota Tempat Tinggal

STATE	Provinsi Tempat Tinggal
Current_Job_Years	Lama Bekerja di Pekerjaan Saat ini
Current_House_Years	Lama Tinggal di Rumah Saat Ini

Tabel 2. Keterangan Atribut Label/Kelas

Variabel	Keterangan
Risk_Flag	Keputusan Peminjaman (label)

### 3. Fase Pengolahan Data

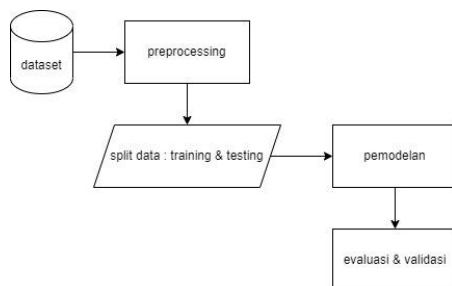
Pada tahap ini dilakukan seleksi atribut, pembersihan data, dan membagi data menjadi *data training* dan *data testing*.

### 4. Fase Pemodelan (Modelling Phase)

Penelitian prediksi ini menggunakan metode SVM dan metode *Naive Bayes* yang akan dibandingkan tingkat akurasi atau kinerjanya. Berikut tahapan keseluruhan pemodelan prediksinya.

### 5. Fase Evaluasi dan Validasi

Pada fase ini dilakukan pengukuran performa model menggunakan teknik *Confusion Matrix*, serta *10-fold Cross Validation* untuk memvalidasi model.

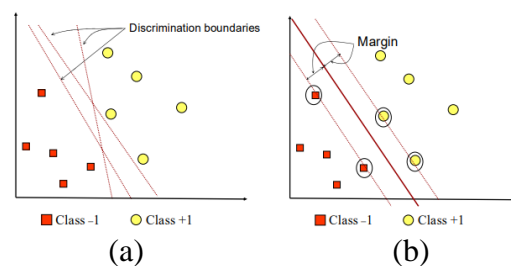


Gambar 2. Tahapan Pemodelan Prediksi **Support Vector Machine**

Teori *Support Vector Machine* (SVM) pertama kali diperkenalkan oleh Vapnik dengan Bernhard Boser dan Isabelle Guyon. Pada dasarnya konsep SVM adalah usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas (Han dan Kamber, 2006).

Algoritma ini digunakan untuk mengklasifikasi data linier dan tidak linier. Untuk mendapatkan *hyperplane* yang baik di antara satu kelas dengan kelas lain dilakukan dengan menghitung lebar *margin* secara maksimal.

Algoritma metode SVM ini dapat mengklasifikasikan data baik itu data linier maupun tidak linier. Untuk mencari *hyperplane* yang optimal di antara satu kelas dengan kelas yang lainnya adalah dengan cara menghitung lebar *margin* secara maksimal. *Margin* adalah jarak dari *hyperplane* atau bidang pemisah optimal terhadap *point* terdekat yang berada di masing-masing kelas. *Point* paling dekat disebut *support vector* (Nugroho, dkk., 2003).



Gambar 3. Menentukan *hyperplane* terbaik antara kelas -1 dan +1 dengan SVM

Gambar 3.(a) dan 3.(b) memperlihatkan beberapa pola anggota dari dua buah kelas -1 dan +1. Pada Gambar 3.(a) *hyperplane* pemisah yang terbaik di antara kedua kelas ditemukan dengan cara mengukur *margin hyperplane* dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* dengan pola terdekat dari setiap kelas. Pola yang paling dekat ini disebut sebagai *support vector*. Pada Gambar 3.(b) menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM.

Diasumsikan bahwa kedua kelas -1 dan +1 dapat terpisah secara sempurna oleh *hyperplane* berdimensi  $d$  yang didefinisikan pada Persamaan (1).

$$w * x_i + b = 0 \quad (1)$$

Pola  $x_i$  yang termasuk kelas 1 dapat dirumuskan sebagai pola yang memenuhi Pertidaksamaan (2).

$$w + x_i + b \geq 1 \quad (2)$$

Sedangkan pola  $x_i$  yang termasuk kelas -1 dirumuskan dengan Pertidaksamaan (3).

$$w * x_i + b \geq -1 \quad (3)$$

Data  $x_i$  sebagai  $x_i \in R^2$ , sedangkan label masing-masing dinotasikan  $y_i \in \{1,0,-1\}$  untuk  $i = 1, 2, 3, \dots, l$  dengan  $l$  adalah banyaknya data. *Margin* terbesar dengan memaksimalkan nilai jarak antara *hyperplane* dan titik terdekat dengan Persamaan (4).

$$\frac{1}{||w||} \quad (4)$$

Hal ini dapat dirumuskan sebagai *quadratic programming problem* yaitu mencari titik minimal Persamaan (5) dengan memperhatikan *constraint* Persamaan (6).

$$\text{Min } \tau(w) = \frac{1}{2} ||w||^2 \quad (5)$$

$$y_i(w, x_i) - 1 \geq 0, \forall_i \quad (6)$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi di antaranya *Lagrange Multipliers*, seperti ditunjukkan pada Persamaan (7).

$$L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^l \alpha_i (y_i((w \cdot x_i) + b) - 1), \quad (7)$$

$$i = 1, 2, \dots, l$$

$\alpha_i$  adalah *Lagrange Multipliers* yang bernilai nol atau positif, yaitu  $\alpha_i \geq 0$ . Nilai optimal dari Persamaan (8) dapat dihitung dengan meminimalkan  $L$  terhadap  $w$  dan  $b$  dan memaksimalkan  $L$  terhadap  $\alpha_i$ . Berdasarkan sifat bahwa pada titik optimal  $L = 0$ , Persamaan (9) dapat dimodifikasi sebagai maksimisasi problem yang hanya mengandung  $\alpha_i$ .

$$\sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j x_i x_j \quad (8)$$

Dari hasil perhitungan diperoleh  $\alpha_i$  yang

kebanyakan bernilai positif. Data yang berkorelasi dengan  $\alpha_i$  yang positif inilah yang disebut *support vector*.

$$\alpha_i \geq 0 \quad (i = 1, 2, \dots, l) \sum_{i=1}^l \alpha_i y_i = 0 \quad (9)$$

### Naive Bayes

*Naive Bayes* merupakan sebuah algoritma pengklasifikasi berbasis probabilitas sederhana, yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan. Teori lain menyebutkan *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya yang dikemukakan oleh Thomas Bayes (Bustami, 2014). Persamaan dari teorema Bayes adalah:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (10)$$

Keterangan:

$X$  : Data dengan kelas yang belum diketahui

$H$  : Hipotesis data  $X$  merupakan suatu kelas spesifik

$P(H|X)$  : Probabilitas hipotesis  $H$  berdasar kondisi  $X$  (*posteriori probability*)

$P(H)$  : Probabilitas hipotesis  $H$  (*prior probability*)

$P(X|H)$  : Probabilitas  $X$  berdasarkan kondisi pada hipotesis  $H$

$P(X)$  : Probabilitas  $X$

### Confusion Matrix

*Confusion Matrix* merupakan pengukuran performa untuk masalah klasifikasi *machine learning* dengan keluaran dapat berupa dua kelas atau lebih. Tabel 3. Tabel *Confusion Matrix*

	Klasifikasi		
		Aktual 0	Aktual 1
Obeservasi	Prediksi 0	True 0	False 1
	Prediksi 1	False 0	True 1
	Total		



$$Akurasi = \frac{True\ 0 + True\ 1}{Prediksi\ 0 + Prediksi\ 1} \times 100\% \quad (11)$$

$$Error\ rate = 100\% - akurasi \quad (12)$$

## HASIL PEMBAHASAN

Proses pelaksanaan penelitian ini dilakukan dengan tahapan berikut.

### Sample Data Set

Pengambilan *dataset* awal menggunakan variabel kriteria yang diambil dari kaggle.com. Atribut yang dipakai sebagai variabel/parameter yang mempengaruhi kelayakan nasabah dalam melakukan peminjaman seperti pada Tabel 1.

Dengan menyesuaikan kondisi di Indonesia maka peneliti menghilangkan atribut *profession*, *city*, *state* karena berbeda dengan karakter nasabah di Indonesia. Dilanjutkan dengan pemilihan atribut yang dijadikan label/kelas sebagai variabel prediksi. Pada Gambar 4 ditampilkan atribut dari kaggle.com.

ExampleSet (Select Attributes)

ExampleSet (Read CSV)

Open

Tutorial Prep

Auto Model

Filter (100/100 examples)

OK

Row No.	ID	Income	Age	Experience	Married	Single	Home_Ownership	Car_Ownership	CURRENT_JOB_YRS	CURRENT_HOUSING_YRS	Risk_Flag
1	1	1205036	23	3	single	rented	no	3	13	0	
2	2	7074616	40	10	single	rented	no	9	13	0	
3	3	3981810	66	4	married	rented	no	4	10	0	
4	4	6205451	41	2	single	rented	yes	2	12	1	
5	5	5708871	47	11	single	rented	no	3	14	1	
6	6	6810837	64	0	single	rented	no	0	12	0	
7	7	3954973	58	14	married	rented	no	6	12	0	
8	8	1708172	33	2	single	rented	no	2	14	0	
9	9	7008040	24	17	single	rented	yes	11	11	0	
10	10	6906460	23	12	single	rented	no	5	13	0	
11	11	4634600	78	7	single	rented	no	7	12	0	
12	12	6602363	22	4	single	rented	no	4	14	0	
13	13	9120880	26	9	single	rented	no	9	12	0	
14	14	8043860	57	12	single	rented	no	8	10	0	
15	15	9420830	48	6	single	rented	no	6	10	1	

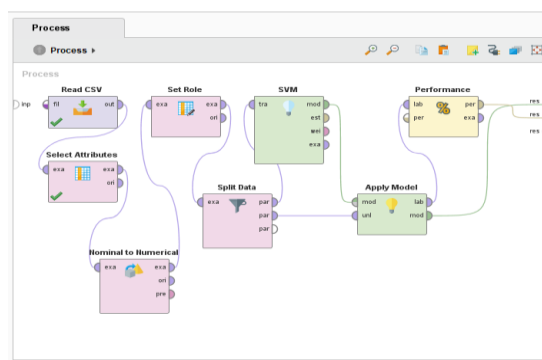
ExampleSet (100 examples, 100 actual attributes, 10 model attributes)

Gambar 4. Data kaggle.com tentang prediksi kredit berisiko

Data tersebut dilakukan *preprocessing data* dengan melakukan proses pada *Rapidminer* dengan menggunakan Operator *Read CSV* untuk membaca data. Atribut yang terpilih diubah tipe datanya untuk atribut yang mempunyai tipe nominal ke tipe numerik dengan menggunakan Operator *nominal to numerical*. Setelah itu dilakukan pembagian menjadi data latih dan data tes

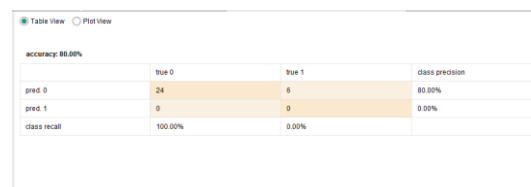
dengan rasio 70% dan 30% dengan operator *Split data*.

Saat data siap diolah maka dengan menggunakan operator SVM yang diberi input data latih, menghasilkan output modul (mod). Kemudian output SVM dijadikan input pada operator *apply model* dan diberi input data tes yang merupakan output dari *split data*. Untuk mengetahui *performance* dari hasil klasifikasi dapat dilihat pada Gambar 5 yang merupakan gambar model proses SVM dalam *Rapidminer*.



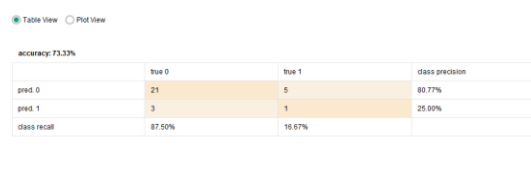
Gambar 5. Model proses SVM dalam *Rapidminer*

Hasil dari *performance* yang dicapai didapatkan tingkat akurasi mencapai 80% dengan nilai perbandingan antara data latih dan data tes adalah 70% dan 30%, dan mencapai tingkat akurasi 80% dengan nilai data perbandingan antara data latih dan data tes adalah 80% dan 20%.



	true 0	true 1	class precision
pred. 0	24	5	80.00%
pred. 1	0	0	0.00%
class recall	100.00%	0.00%	

Gambar 6. Hasil tingkat akurasi dengan SVM



	true 0	true 1	class precision
pred. 0	21	5	80.77%
pred. 1	3	1	25.00%
class recall	87.50%	16.67%	

Gambar 7. Hasil tingkat akurasi dengan

### *Naive Bayes*

## **KESIMPULAN DAN SARAN**

Penelitian ini menemukan tingkat akurasi yang dihasilkan oleh Algoritma SVM lebih tinggi, yaitu 80%, dibandingkan dengan Algoritma *Naive Bayes* 73.33%. Hasil pengukuran performa model menunjukkan *true positive* sebanyak 24 nasabah dari 30 nasabah. Model prediksi yang diusulkan telah tervalidasi sehingga dapat dimanfaatkan untuk bahan pengetahuan pengambilan kebijakan.

Saran pengembangan penelitian ke depan, model prediksi dapat diimplementasikan dan divalidasi di sektor selain perbankan menggunakan prediktor yang berbeda.

## **DAFTAR PUSTAKA**

- Azizah, N. 2022. *No Title*. <https://www.republika.co.id/berita/r7cm10463/46-persen-perempuan-pemilik-umkm-kesulitan-bayar-tagihan-dan-utang>.
- Boswell, D. 2003. *An Introduction to Support Vector Machines*. Recent Advances and Trends in Nonparametric Statistics. <https://doi.org/10.1016/B978-044451378-6/50001-6>.
- Bustami. 2014. Penerapan Algoritma *Naive Bayes* Untuk Nasabah Asuransi. *Jurnal Informatika*, 8(1), 884–898.
- Han, J. dan Kamber, M. 2006. Data mining: Data mining concepts and techniques. In Asma Stephan (Ed.), *Morgan Kaufmann Publishers is an imprint of Elsevier* (Second Edi). Diane Cerra. [www.mkp.com](http://www.mkp.com) or [www.books.elsevier.com](http://www.books.elsevier.com)
- Iskandar, J.W. dan Nataliani, Y. 2021. Perbandingan *Naive Bayes*, SVM, dan k-NN untuk Analisis Sentimen Gadget Berbasis Aspek. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 5(6), 1120–1126. <https://doi.org/10.29207/resti.v5i6.3588>
- Nugroho, A.S., Witarto, A.B., dan Handoko, D. 2003. *Support Vector Machine –Teori dan Aplikasinya dalam Bioinformatika –*. <http://ci.nii.ac.jp/naid/110002935335/>
- Pertiwi, M.W. 2019. Analisis Sentimen Opini Publik Mengenai Sarana dan Transportasi Mudik Tahun 2019 Pada Twitter Menggunakan Algoritma *Naive Bayes*, Neural Network, K-NN dan SVM. *Inti Nusa Mandiri*, 14(1), 27–32.
- Zainuddin, N. dan Selamat, A. 2014. Sentiment analysis using Support Vector Machine. *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings, September*, 333–337. <https://doi.org/10.1109/I4CT.2014.6914200>