Indones. J. Math. Nat. Sci. 46(1) 2023
Indonesian Journal of Mathematics and Natural Sciences
http://journal.unnes.ac.id/nju/index.php/JM

Komparasi Metode Support Vector Machine dan Naive Bayes Classifier untuk Pemodelan Kualitas Pengajuan Kredit

Moh Khubaib Tamami*, Iqbal Kharisudin

Jurusan Matematika, FMIPA, Universitas Negeri Semarang, Indonesia
Gedung D7 Lt.1, Kampus Sekaran Gunungpati, Semarang 50229
E-mail: khbbtmm@gmail.com

Diterima 7 Desember 2022 Disetujui 10 Februari 2023 Dipublikasikan 28 April 2023

Abstrak

Kredit bermasalah atau Non-Performing Loan (NPL) merupakan suatu risiko yang terdapat dalam setiap pemberian kredit oleh bank atau institusi pemberian kredit kepada customernya. Suatu bank memiliki risiko yang tinggi dalam hal kredit dapat dilihat dari tinggi rendahnya NPL bank tersebut. Permasalahan NPL ditandai dengan kredit yang tidak dapat kembali tepat pada jangka waktu yang telah ditentukan. Analisis penilaian risiko calon debitur pada suatu bank perlu dilakukan untuk mengantisipasi permasalahan NPL. Pada penelitian ini dilakukan analisis pemodelan kualitas kredit menggunakan metode Support Vector Machine (SVM) dan Naive Bayes Classifier (NBC) untuk melakukan klasifikasi penilaian risiko calon debitur. Pada SVM diujikan tiga fungsi kernel, yaitu fungsi kernel Linear, fungsi kernel Gaussian Radial Basis Function (RBF), dan fungsi kernel Polynomial. Hasil penelitian menunjukkan model klasifikasi yang dihasilkan oleh metode SVM fungsi kernel RBF merupakan model terbaik, dengan akurasi 96.65%. Hasil pemodelan dapat digunakan dalam mengklasifikasi sekaligus menyeleksi calon debitur, apakah akan menjadi debitur yang mempunyai kualitas kredit baik atau buruk.

Kata kunci: kredit bermasalah, kredit, *data mining*, SVM, NBC

Abstract

Non-performing loans (NPL) are a risk inherent in every credit extension by bank or credit institution to its customers. A bank has a high risk in terms of credit, which can be seen from the high and low NPL of the bank. NPL problems are characterized by credit that cannot be returned at a predetermined time. Analysis of the risk assessment of potential debtors at a bank is necessary to anticipate NPL problems. In this study, an analysis of credit quality modeling was carried out using the Support Vector Machine (SVM) and Naive Bayes Classifier (NBC) method to classify the risk assessment of potential debtors. In SVM, three kernel functions were tested, they are the Linear kernel function, the Gaussian Radial Basis Function (RBF) kernel function, and the Polynomial kernel function. The results showed that the classification model produced by the SVM method of the RBF kernel function is the best model, with an accuracy of 96.65%. Modeling results can be used in classifying as well as selecting prospective debtors, whether to be debtors with good or bad credit quality.

Keywords: *non-performing loan*, *credit, data mining*, SVM, NBC.

How to cite:
Tamami, M. K., Kharisudin, I. (2023). Komparasi metode support vector machine dan naive bayes classifier untuk pemodelan kualitas pengajuan kredit. *Indonesian Journal of Mathematics and Natural Sciences*, 46(1), 38-44.

PENDAHULUAN

Kredit bermasalah atau *Non-Performing Loan* (NPL) merupakan suatu risiko yang terdapat dalam setiap pemberian kredit. NPL juga dapat dikatakan sebagai pinjaman yang belum dibayar (Khan *et al*., 2019). Suatu bank memiliki risiko yang tinggi dalam hal kredit dilihat dari tinggi rendahnya NPL bank tersebut. Permasalahan NPL ditandai dengan kredit yang tidak dapat kembali tepat pada jangka waktu yang telah ditentukan, sehingga dapat dikatakan bahwa NPL menjadi indikasi adanya masalah dalam bank. Bank akan mengalami kerugian yang besar jika masalah NPL tidak segera mendapatkan

solusi (Firdaus, 2017). Salah satu penanganan yang dilakukan adalah dengan mengadakan analisis risiko kredit. Pada pelaksanaannya, seorang analis kredit perbankan perlu mengambil keputusan yang tepat, apakah akan menerima atau menolak pengajuan kredit oleh nasabah.

Pada penelitian ini dilakukan analisis pemodelan kualitas pengajuan kredit dengan menerapkan algoritma data mining. Data mining merupakan proses mengekstraksi serta mengidentifikasi pengetahuan yang bermanfaat yang dapat digunakan di masa mendatang dari data yang masif dengan menggunakan matematika, statistika, *artificial intelligence* (AI) dan *machine learning* (Hermawati, 2013). Pemanfaatan dari data mining salah satunya adalah klasifikasi. Klasifikasi sendiri memiliki definisi yaitu pengelompokkan data yang memiliki kelas label atau target. Berbagai metode klasifikasi mengalami kemajuan, salah satunya di bidang finansial. Bank sebagai lembaga yang bergerak dalam hal finansial memiliki peranan salah satunya yaitu memberikan kartu kredit kepada nasabah. Namun, permasalahan yang muncul adalah sulitnya menganalisis risiko kredit.

Penelitian mengenai aplikasi klasifikasi NBC untuk menentukan pengajuan, kartu kredit oleh nasabah di BNI Syariah Surabaya (Antaristi & Kurniawan, 2017), metode NBC dimanfaatkan untuk mencari pola nasabah, apakah nasabah diterima atau tidak, menggunakan beberapa variabel yakni jenis kelamin, status, banyak tanggungan, profesi, status rumah dan penghasilan per tahunnya. Didapatkan rata-rata nilai *recall* 95%, nilai *precission* 100%, dan nilai *accuracy* 98.67%.

Penelitian lain dilakukan oleh Rifqo mengenai implementasi algoritma NBC dalam penentuan pemberian kredit (Rifqo & Wijaya, 2017). Tujuan dari penelitian ini adalah untuk mendapatkan tingkat akurasi dalam menganalisa kredit mobil. Data yang digunakan adalah data yang didapatkan dari UCI *data set,* mengenai persutujuan kredit antara Japan dan Australia, kemudian data kredit (Agiing) perusahaan *leasing* ACC tahun 2010 sampai dengan 2011. Penelitian menunjukkan hasil bahwa NBC memiliki akurasi yang baik dalam memprediksi *approval,* yaitu Agiing 2010 *accuracy* 96.59%, Agiing 2011 *accuracy* 97.59%. *Japan credit approval accuracy* 74.89%, *Australia credit approval accuracy* 80%.

Penelitian terkait NLP yang pernah dilakukan yaitu pengujian regresi dengan memanfaatkan teknik data mining regresi linier berganda dan teknik klasifikasi *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) (Bisht *et al*., 2020). Diperoleh hasil bahwa SVM menjadi metode terbaik karena memiliki akurasi tertinggi. Akurasi SVM yang tinggi didukung pula pada penelitian yang membandingkan metode SVM, Regresi Logistik, dan Analisis Diskriminan. Hasilnya, SVM dapat menentukan fitur-fitur kegagalan pembayaran kredit lebih banyak (Bellotti & Crook, 2009). Selaras pula dengan penelitian lain yang menunjukkan bahwa SVM (SVM Kernel Non-Linier) unggul secara statistik sebesar 95% dibanding Regresi Logistik (Haltuf, 2014). Kemudian penelitian yang mengkombinasikan SVM dengan *Metaheuristic Approaches* (MA). Hasilnya yaitu SVM lebih handal dalam mendeteksi fitur-fitur NLP. Akan tetapi kombinasi antara SVM dan MA dapat membentuk sebuah model perhitungan yang transparan dan kompetitif (Goh & Lee, 2019)

Penelitian lain juga dilakukan dengan mengkombinasikan *Data Envelopment Analysis* (DEA) sebagai metode pembobotan dengan DMADM sebagai metode pengambilan keputusan. Kemudian hasilnya dirangking dengan metode *Correlation Coefisient and Standart Deviation* (CCSD). Hasilnya yaitu CCSD memiliki konsistensi yang tinggi dalam perhitungan. Sehingga metode ini dapat digunakan untuk data dengan interval waktu yang berbeda-beda. Kemudian metode DEA lebih handal dalam kriteria pembobotan data. Terlebih metode ini mengkombinasikan 5 metode MADM yang mampu meminimalisir kekurangan metode lain dalam pengambilan keputusan (Dahooie *et al*., 2021). Disisi lain, peneliti juga mengintegrasikan pengetahuan pakar dengan *Genetic Algorithms* (GA). Hasilnya yaitu kombinasi antara GA dengan kNN serta GA dengan *Naïve Bayes* (NB) berkinerja paling baik. Terlebih dengan adanya pengetahuan pakar mampu membuat solusi yang lebih optimal (Lappas & Yannacopoulos, 2021).

METODE

Data yang digunakan dalam penelitian ini adalah data status transaksi kartu kredit nasabah yang diunduh dari website www.github.com. Data terdiri dari 1 variabel terikat (Y) yaitu status transaksi kartu kredit nasabah dengan kategori lancar (0) dan tidak lancar (1), dengan 10 variabel bebas (X), antara lain X_1 merupakan jumlah pokok utang pinjaman yang tidak memiliki jaminan dan pokok utangnya bisa naik turun sesuai penggunaan dibagi dengan jumlah plafon nasabah, X_2 merupakan umur nasabah, X_3 merupakan Frekuensi nasabah pernah terlambat melakukan pembayaran 30-59 hari dalam kurun waktu dua tahun , X_4 merupakan *debt ratio*, X_5 merupakan pendapatan bulanan nasabah, X_6 merupakan jumlah fasilitas kredit yang aktif, X_7 merupakan frekuensi nasabah pernah

terlambat melakukan pembayaran 90 hari/lebih, X_8 merupakan jumlah fasilitas kredit rumah, X_9 merupakan frekuensi nasabah pernah terlambat melakukan pembayaran 60-89 hari dalam kurun waktu dua tahun, X_{10} merupakan jumlah tanggungan keluarga (kecuali debitur). Jumlah data ada 872. Data tersebut akan dianalisis menggunakan metode SVM dengan perbandingan 3 fungsi kernel, dan metode NBC.

Selanjutnya setelah data diperoleh kemudian dilakukan tahap *preprocessing* data. Hal pertama yang dilakukan dalam *preprocessing* data adalah standardisasi, yaitu untuk menyamakan skala pada data. Pada data yang digunakan. antar data memiliki skala yang cukup tinggi, sehingga diperlukan standardisasi. Setelah data memiliki skala data yang sama, dilakukan uji outlier, yaitu untuk mengurangi bias pada model klasifikasi. Kemudian dilakukan seleksi data, yaitu mengkorelasikan antar variabel independen. Hal ini dilakukan supaya antar variabel independen tidak memiliki korelasi yang kuat sehingga model yang didapatkan akan tepat dan optimal. Setelah melewati tahap *preprocessing,* dilakukan pembagian data menjadi data *training* dan data *testing.*

Kemudian dilakukan tahap klasifikasi menggunakan 2 algoritma. Algoritma pertama yang digunakan adalah *Support Vector Machine* (SVM) dengan perbandingan fungsi kernel Linear, *Gaussian Radial Basic Function* (RBF), dan Polynomial. Algoritma SVM memiliki tujuan untuk menemukan *hyperplane* yang paling optimal dengan margin maksimum yang bertindak untuk memisahkan dua kelas yang berbeda (Goh & Lee, 2019). Margin sendiri merupakan jarak antara *hyperplane* dengan data terdekat dari masing-masing kelas.

Algoritma kedua yang digunakan adalah *Naive Bayes Classifier* (NBC). NBC sendiri merupakan pengklasifikasi probabilistik sederhana yang menghitung sekumpulan peluang dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan (Prasetyo, 2012). NBC ini didasarkan pada asumsi penyederhanaan bahwa nilai variabel secara kondisional saling bebas jika diberikan nilai *output.* Dengan kata lain, jika diberikan *output,* peluang mengamati secara bersamaan adalah produk dari peluang individu (Ridwan *et al*., 2013).

Kedua algoritma tersebut akan dibandingkan dan dilihat performa kinerjanya dalam mengklasifikasi data dengan evaluasi model k-*fold cross validation.* Penggunaan k-*fold cross validation* bertujuan untuk menghilangkan bias pada data (Bramer, 2007).

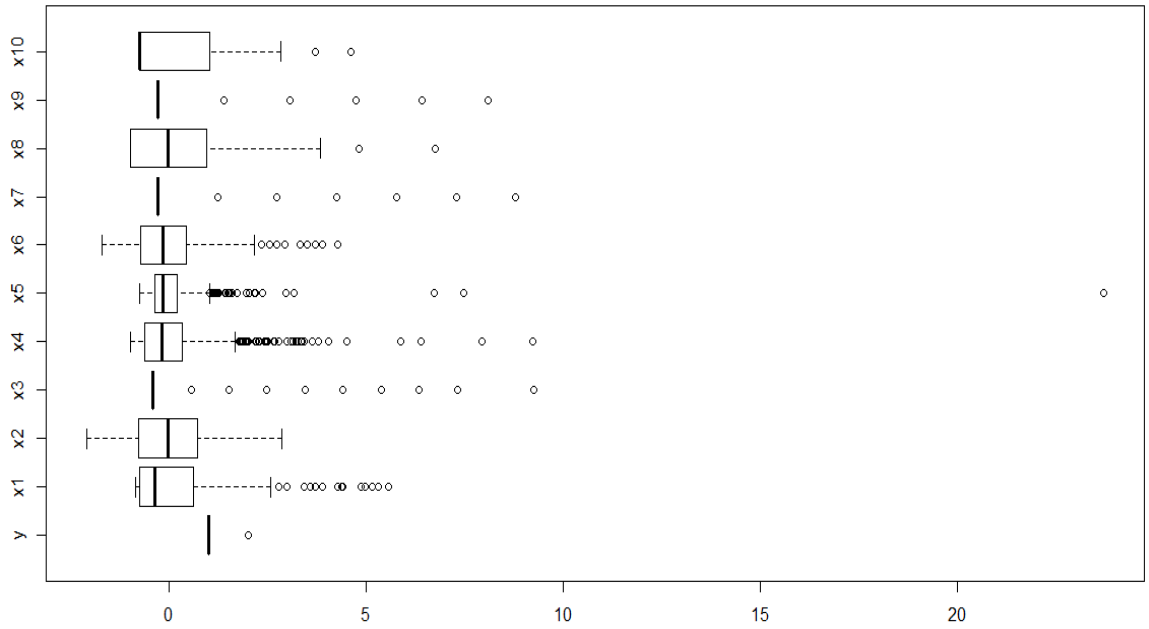
HASIL DAN PEMBAHASAN

Data yang telah diperoleh kemudian dilakukan tahap *preprocessing.* Hal yang pertama dilakukan adalah standardisasi. Standardisasi bertujuan untuk menyamakan skala pada data. Hasil dari tahap ini dapat dilihat pada Gambar 1.

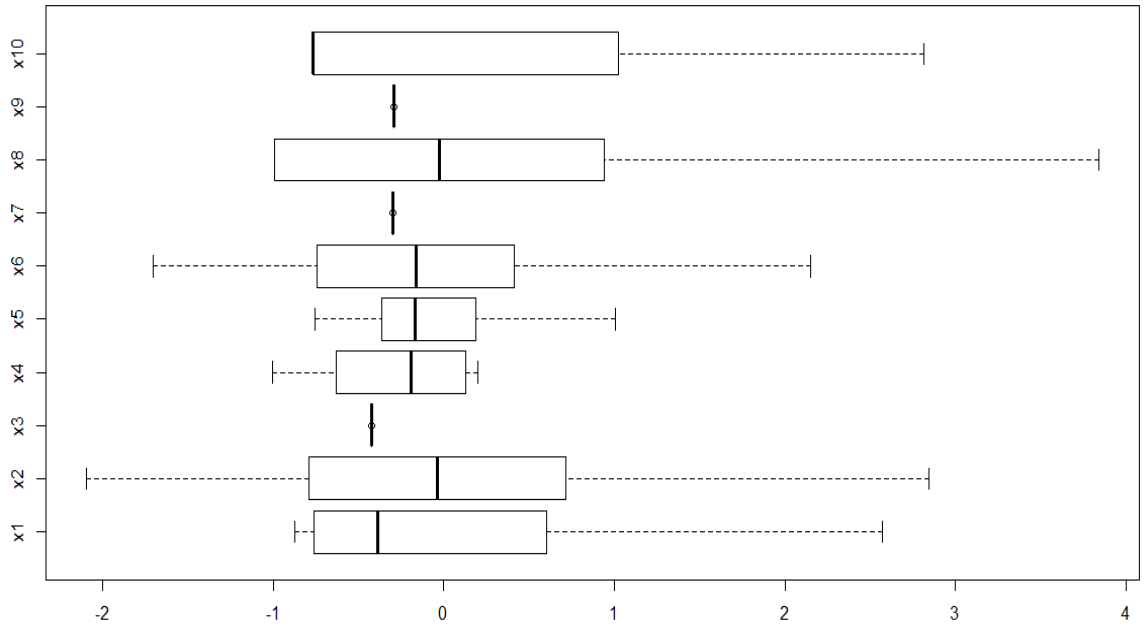
```
y            x1                  x2                  x3                  x4
0:750   Min.   :-0.8714   Min.   :-2.09414   Min.   :-0.4207   Min.   :-1.0042
1:122   1st Qu.:-0.7633   1st Qu.:-0.79126   1st Qu.:-0.4207   1st Qu.:-0.6298
        Median :-0.3850   Median :-0.03696   Median :-0.4207   Median :-0.1942
        Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
        3rd Qu.: 0.6001   3rd Qu.: 0.71734   3rd Qu.:-0.4207   3rd Qu.: 0.3112
        Max.   : 5.5547   Max.   : 2.84309   Max.   : 9.2340   Max.   : 9.2184
      x5                  x6                  x7                  x8
Min.   :-0.7539   Min.   :-1.7055   Min.   :-0.2971   Min.   :-0.99368
1st Qu.:-0.3664   1st Qu.:-0.7414   1st Qu.:-0.2971   1st Qu.:-0.99368
Median :-0.1667   Median :-0.1630   Median :-0.2971   Median :-0.02662
Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
3rd Qu.: 0.1856   3rd Qu.: 0.4155   3rd Qu.:-0.2971   3rd Qu.: 0.94045
Max.   :23.7133   Max.   : 4.2718   Max.   : 8.7931   Max.   : 6.74283
      x9                 x10
Min.   :-0.294   Min.   :-0.7694
1st Qu.:-0.294   1st Qu.:-0.7694
Median :-0.294   Median :-0.7694
Mean   : 0.000   Mean   : 0.0000
3rd Qu.:-0.294   3rd Qu.: 1.0220
Max.   : 8.085   Max.   : 4.6049
```

Gambar 1. Hasil standardisasi

Setelah dilakukan standardisasi, kemudian dilakukan tahap uji outlier, yaitu bertujuan untuk mengurangi bias pada model klasifikasi. Proses ini dilakukan dengan melihat *boxplot* pada masing-masing variabel seperti pada Gambar 2. Karena masih banyak *outlier* pada data, sehingga dilakukan metode IQR. Hasil dari *boxplot* menggunakan metode IQR dapat dilihat pada Gambar 3.

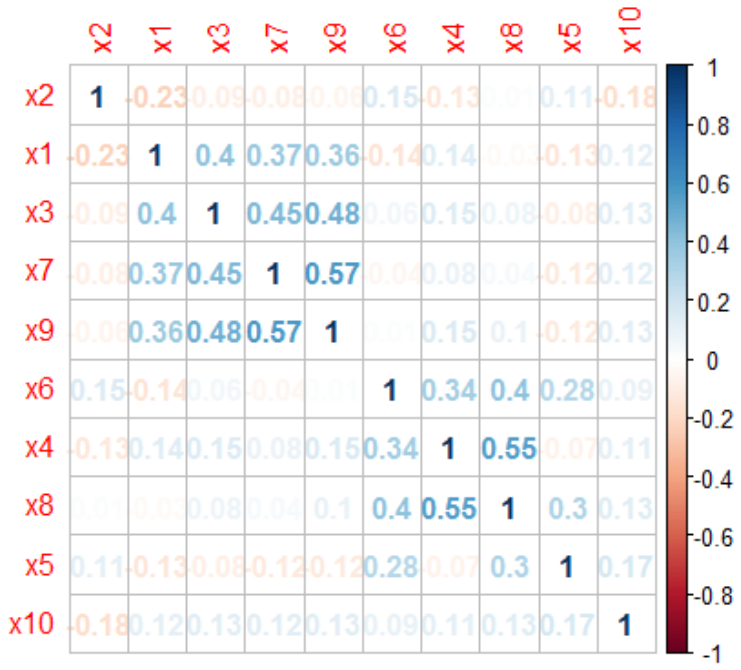


Gambar 2. *Boxplot* awal semua variabel



Gambar 3. *Boxplot* dengan Metode IQR

Berdasarkan hasil dari *imputation* menggunakan metode IQR, terlihat bahwa sudah tidak ada data yang *outlier.* Kemudian dilakukan seleksi data. Tahapan ini dilakukan dengan mengkorelasikan antar variabel independen, supaya antar variabel independen tidak memiliki korelasi yang kuat sehingga model yang didapatkan akan tepat dan optimal. Dari semua variabel independen yang dikorelasikan, diperoleh seperti pada Gambar 4.



Gambar 4. Matrix korelasi seleksi data

Data yang sudah melalui tahap *preprocessing,* kemudian dilakukan tahap pembagian data menjadi data *training* dan data *testing.* Perbandingan yang digunakan adalah 80:20, 80% untuk data *training* dan 20% untuk data *testing*. Adapun uraiannya dijelaskan pada Tabel 1.

Tabel 1. Pembagian data *training* dan *testing*

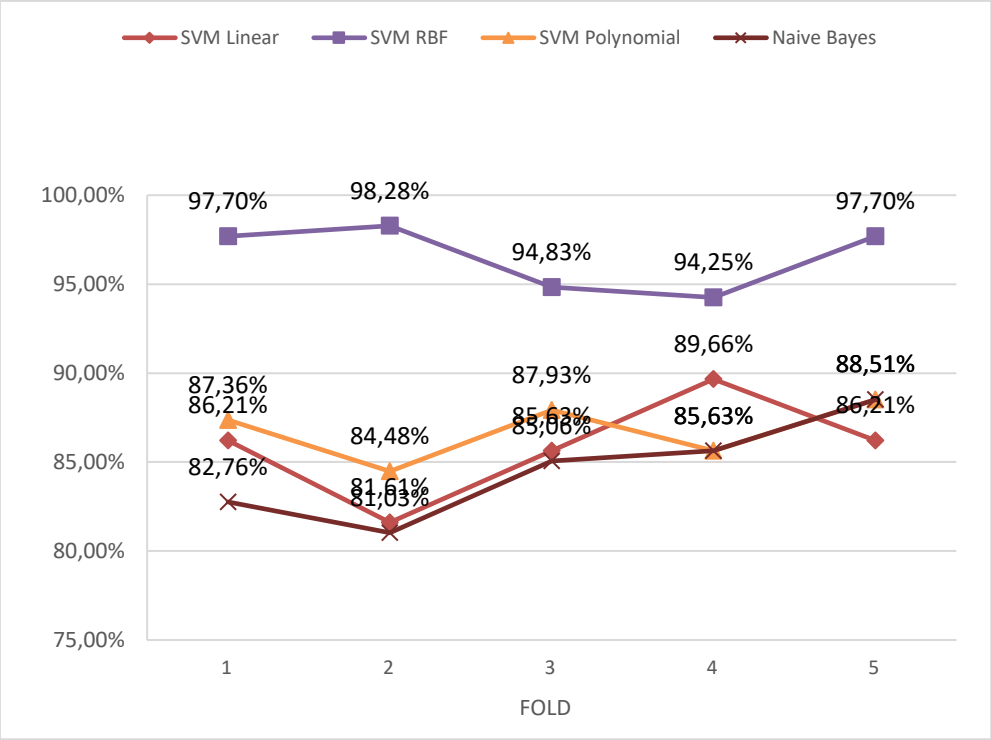
Data	Lancar	Tidak Lancar	Jumlah
Training	596	92	688
Testing	154	30	184
Jumlah	750	122	872

Dari hasil pembagian data *training* dan *testing* tersebut, dapat diketahui bahwa adanya *imbalancing* data, yaitu jumlah data lancar jauh lebih banyak dibandingkan dengan data tidak lancar. Oleh karena itu, perlu dilakukan penyeimbangan data dengan metode *over and under sampling*, yaitu teknik penyeimbangan data dengan mengurangi jumlah kelas lancar dan menambah jumlah kelas tidak lancar.

Kemudian dilakukan klasifikasi menggunakan SVM dan NBC. Untuk SVM sendiri digunakan tiga fungsi kernel, yaitu fungsi kernel Linear, fungsi kernel RBF, dan fungsi kernel Polynomial. Perbandingan kinerja algoritma dapat dilihat pada Gambar 5.

Algoritma pertama yang digunakan untuk mengklasifikasi data status transaksi kartu kredit nasabah adalah SVM, dengan tiga fungsi kernel, yaitu fungsi kernel Linear, fungsi kernel RBF, dan fungsi kernel Polynomial. Setelah dilakukan klasifikasi dengan evaluasi model 5-*fold cross validation* diperoleh rata-rata *accuracy* sebesar 85.86% untuk SVM fungsi kernel Linear dengan parameter $\boldsymbol{C = 1}$, 96.55% untuk SVM fungsi kernel RBF dengan parameter $\boldsymbol{C = 5}$ dan $\boldsymbol{\gamma = 2}$, dan 86.78% untuk SVM fungsi kernel Polynomial dengan parameter $\boldsymbol{C = 100}$ dan $\boldsymbol{d = 2}$. Hasil tersebut merupakan hasil yang tinggi, hal ini dikarenakan SVM memiliki kelebihan dapat mengolah data berdimensi tinggi, tanpa mengalami penurunan performa (Purnawaman, 2015).

Algoritma kedua yang digunakan adalah NBC. Setelah dilakukan klasifikasi dengan evaluasi model 5-*fold cross validation* diperoleh rata-rata *accuracy* sebesar 85.59%. Nilai *accuracy* ini terbilang cukup tinggi, dikarenakan NBC memiliki kelebihan yaitu tergolong algoritma yang sederhana serta memiliki kompleksitas perhitungan yang rendah, namun memiliki kelemahan dimana sifat independensi dari fitur NBC tidak dapat selalu diterapkan sehingga nantinya akan mempengaruhi tingkat akurasi perhitungan (Pamungkas, 2020).



Gambar 5. Diagram perbandingan *accuracy* algoritma

Dari kedua algoritma tersebut, maka dapat diketahui bahwa SVM dengan fungsi kernel RBF merupakan algoritma paling cocok untuk mengklasifikasikan data status transaksi kartu kredit nasabah dibandingkan dengan SVM fungsi kernel RBF, SVM fungsi kernel Polynomial, dan NBC.

SIMPULAN

Berdasarkan hasil dalam pembahasan di atas, maka dapat disimpulkan bahwa hasil klasifikasi data status transaksi kartu kredit menggunakan algoritma SVM fungsi kernel Linear, SVM fungsi kernel RBF, dan SVM fungsi kernel Polynomial dengan evaluasi model 5-*fold cross validation* berturut-turut menghasilkan *accuracy* sebesar 85.86%, 96.55%, dan 86.78%. Kemudian klasifikasi data transaksi kartu kredit menggunakan algoritma NBC dengan evaluasi model 5-*fold cross validation* menghasilkan *accuracy* sebesar 85.59%. Jadi, dapat disimpulkan bahwa algoritma SVM fungsi kernel RBF merupakan algoritma yang paling cocok untuk mengklasifikasikan data status transaksi kartu kredit dibandingkan dengan SVM fungsi kernel Linear, SVM fungsi kernel Polynomial, dan NBC.

DAFTAR PUSTAKA

Antaristi, M., & Kurniawan, Y. I. (2017). Aplikasi klasifikasi penentuan pengajuan kartu kredit menggunakan metode naive bayes di Bank BNI Syariah Surabaya. *Jurnal Teknik Elektro*, 9(2), 45-52.

Bellotti, T., & Crook, J. (2009). Support vector machine for credit scoring and discovery of significant features. *Expert Systems with Application*, 36(2009), 3302-3308.

Bisht, R., Rathore, T., & Dixit, P. (2020). Non performing loans financial risk: a study and analytics though data mining. *Journal of Critical Reviews,* 7(15), 3905-3915.

Bramer, M. (2007). *Principles of Data Mining*. New York: Springer.

Dahooie, J. H., Hajiagha, S. H. R., Farazmehr, S., Zavadskas, E. K., & Antucheviciene, J. (2021). A novel dynamic credit risk evaluation method using data envelopment analysis with common weights and combination of multi-attribute decision-making methods. *Computers and Operational Research*, 129(2021), 1-18.

Firdaus, N. N. (2017). Analisis determinan non performing loan pada bank umum konvensional di Indonesia. *Skripsi*. Fakultas Ekonomi Universitas Negeri Yogyakarta, Yogyakarta.

Goh, R. Y., & Lee, L. S. (2019). Credit scoring: a review on support vector machines and metaheuristic

approaches. *Hindawi: Advances in Operations Research*, 2019, 1-31. https://doi.org/10.1155/2019/1974794.
Haltuf, B. M. (2014). Support vector machine for credit scoring. *Skripsi*. Faculty of Finance University of Economic in Prague, Prague.
Hermawati, F. A. (2013). *Data Mining*. Andi Offset: Yogyakarta.
Khan, M.A., Siddique, A., & Sarwar, Z. (2019). Determinants of non-performing loans in the banking sector in developing state. *Asian Jouenal of Accounting,* 5(1), 135-145.
Lappas, P. Z., & Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk asessment. *Applied Soft Computing Journal*, 107(2021), 1-23.
Pamungkas, F. S. (2020). Analisis Sentimen dengan SVM, NAÏVE BAYES, dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi COVID-19 Pada Media Sosial Twitter. PRISMA, *Prosiding Seminar Nasional Matematika*. Semarang 17 Oktober 2020.
Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Andi Offset Yogyakarta.
Purnawaman, I. (2015). Support vector machine pada information retrieval. *Jurnal Pendidikan Teknologi Dan Kejuruan*, 12(5), 173-180.
Ridwan, M., Suyono, H. & Sarosa, M. (2013). Penerapan data mining untuk evaluasi kinerja akademik mahasiswa menggunakan algoritma naive bayes classifier. *Jurnal EECCIS*, 7(1), 59-64.
Rifqo, M. H., & Wijaya, A. (2017). Implementasi Algoritma naive bayes dalam penentuan pemberian kredit. *Jurnal Pseudocode*, 4(2), 120-128.