Perbandingan 4 Algoritma Klasifikasi untuk Prediksi Resiko Kredit

Devina Adinda Hartono[1]

Program Studi S1 Informatika, Fakultas Informatika Telkom University
Bandung, Indonesia

[1]devinaadinda@student.telkomuniversity.ac.id

Abstract— **Resiko kredit adalah suatu resiko kerugian ketika seorang peminjam tidak mampu membayar hutangnya. Klasifikasi data mining dapat membantu analisis kredit agar tidak terjadinya resiko gagal bayar oleh peminjam. Algoritma klasifikasi antara lain:** *Random Forest, Support Vector Machine, Linear Regression, Gradient Boost***. Diperlukan evaluasi menggunakan *confusion matrix, ROC curve,* nilai AUC, akurasi dan kompleksitas waktu untuk mengetahui algoritma mana yang paling akurat untuk prediksi resiko kredit.**

Kata Kunci — **Resiko Kredit*, Random Forest, Support Vector Machine, Linear Regression, Gradient Boost***

I. PENDAHULUAN

Penelitian mengenai resiko kredit telah banyak dilakukan dan menjadi salah satu masalah yang menarik dalam analisa keuangan. Salah satu solusi yang bisa dilakukan adalah dengan menfaatkan *data mining* untuk melakukan prediksi resiko kredit. Karena *data mining* memiliki mekanisme pembelajaran mandiri setelah dilakukan suatu proses pelatihan *(training)*[1].

Resiko kredit adalah suatu resiko kerugian ketika seorang peminjam tidak mampu membayar hutangnya, baik hutang pokok maupun bunganya. Tingkat resiko kredit berpengaruh terhadap resiko gagal bayar oleh peminjam.

Berdasarkan penelitian *data mining* tentang resiko kredit yang dilakukan pada tahun 2000 sampai 2010 menunjukan algoritma *Support Vector Machine* (SVM) merupakan teknik yang banyak diusulkan oleh para peneliti [2]. Namun pada penelitian Jozef Zurada tahun 2011 menunjukan performa SVM belum bisa mengungguli LR, NN, kNN pada *Quinlan data set.*

Metode *ensamble* pada *machine learning* yang dapat digunakan untuk meningkatkan akurasi suatu klasifikasi adalah *Bagging* dan *Boosting*. Random Forest menerapkan metode *bootstrap aggregating (bagging*) dan *random feature selection*[3]. Dalam penggunaannya, Random Forest dapat menghindari o*verfitting* pada sebuah dataset saat mencapai akurasi maksimum namun terdapat ketidakseimbangan kelas yang merupakan salah satu permasalahan pada *data mining.* Penggunaan *Boosting* berfokus pada *misclassified tuples* dan memiliki kecenderungan peningkatan akurasi yang lebih tinggi dibandingkan *Bagging.* Salah satu algoritma boosting adalah *Gradient Boosting.*

Pada penelitian ini, akan dilakukan perbandingan menggunakan algoritma *Support Vector Machine* (SVM), *Logistic Regression* (LR), *Random Forest* (RF) dan *Gradient Boosting* (GB) untuk prediksi resiko kredit pada German Credit Data.

II. LANDASAN TEORI

A. Data Mining

Data mining merupakan suatu cara dalam penggalian informasi dari sejumlah data yang biasanya tersimpan dalam repositori dengan menggunakan teknologi pengenalan pola, statistik dan teknik matematika.

Data mining merupakan bagian dari proses *Knowledge Discovery in Database* (KDD). Proses KDD terdiri dari langkah-langkah berikut [4]:

a. *Data cleaning,* menghilangkan noise dan data yang tidak konsisten
b. *Data integration,* mengintegrasi beberapa sumber data yang dapat digabungkan.
c. *Data selection,* menyeleksi data yang relevan dengan analisis yang diambil dari *database*.
d. *Data transformation,* proses data ditransformasikan ke dalam format yang sesuai untuk proses dalam *data mining.*
e. *Data mining,* proses esensial dimana metode diaplikasikan untuk mengekstrak pola data.
f. *Pattern evaluation,* proses untuk mengindentifikasikan pola-pola yang menarik untuk dipresentasikan ke dalam *knowledge based*
g. *Knowledge presetation,* proses visualisasi dan teknik representasi yang digunakan untuk menyajikan pengetahuan kepada pengguna.

B. Metode

Metode yang digunakan untuk mengukur akurasi algoritma klasifikasi antara lain [5]:

a. *Confusion Matrix*

Confusion matrix adalah tabel yang mencatat hasil kerja klasifikasi. Kelas yang di prediksi ditampilkan dibagian atas matrix dan kelas data aktual dibagian kiri. Setiap sel berisi

angka yang menunjukan berapa banyak kasus yang sebenarnya dari kelas yang diamati untuk diprediksi.

Tabel 1. Model Confusion Matrix (Bramer, 2007)

Nilai prediksi	Nilai Aktual	
	+	-
+	TP	FN
-	FP	TN

TP : jumlah data positif yang terklasifikasi dengan benar.(Good/1)

TN : jumlah data negatif yang terklasifikasi dengan benar.(Bad/1)

FP : jumlah data positif namun tidak terklasifikasi. (Good/0)

FN : jumlah data negatif namun tidak terklasifikasi (Bad/0)

Setelah data *test* dimasukan kedalam *confusion matrix*, dilakukan perhitungan *accuracy*, *precision*, *recall* dan *f-measure* dengan persamaan sebagai berikut[6]:

1. *Accuracy*

Accuracy (akurasi) adalah jumlah perbandingan data yang benar dengan jumlah keseluruhan data.

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

2. *Precision*

Precision digunakan untuk mengukur seberapa besar proporsi dari kelas data positifyang berhasil diprediksi dengan benar dari keseluruhan hasil prediksi kelas positif.

$$Precision = \frac{TP}{TP + FP}$$

3. *Recall*

Recall digunakan untuk menunjukan kelas data positif yang berhasil diprediksi dengan benar dari keseluruhan data.

$$Recall = \frac{TP}{TP + FN}$$

4. *F-measure*

F-measure merupakan gabungan dari *precision* dan *recall* yang digunakan untuk mengukur kemampuan algoritma dalam mengklasifikasi kelas minoritas.

$$F\text{-}measure = \frac{2\ x\ precision\ x\ \ recall}{precision + recall}$$

Semakin tinggi hasilnya maka algoritma yang dihasilkan dengan metode tersebut semakin baik dalam melakukan klasifikasi.

b. Kurva ROC *(Receiver Operating Characteristic)*

Kurva ROC adalah grafik 2 dimensi dengan *false positives* sebagai garis horizontal (sumbu x) dan *true positives* sebagai garis vertical (sumbu y). Kurva ROC menunjukan akurasi dan membandingkan klasifikasi secara visual. Metode yang digunakan untuk menghitung luas daerah dibawah ROC yang disebut AUC (Area Under Curve) untuk menentukan klasifikasi mana yang lebih baik. Semakin besar AUC maka semakin baik klasifikasi yang digunakan. AUC sangat berguna ketika datasetnya sangat *unbalance* dan kurva ROC sangat berguna ketika probabilitas prediksi tidak "properly calibrated:.

Untuk klasifikasi data mining, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011):

- 0,90 - 1,00 = klasifikasi sangat baik
- 0,80 - 0,90 = klasifikasi baik
- 0,70 - 0,80 = klasifikasi cukup
- 0,60 - 0,70 = klasifikasi buruk
- 0,50 - 0,60 = klasifikasi salah

c. Algoritma Random Forest (RF)

Random Forest menerapkan metode *bootstrap aggregating (bagging*) dan *random feature selection.* Dalam penggunaannya, Random forest dapat menghindari o*verfitting* pada sebuah dataset saat mencapai akurasi maksimum namun terdapat ketidakseimbangan kelas yang merupakan salah satu permasalahan pada *data mining.*

d. Algoritma Support Vector Machine (SVM)

Support Vector Machine merupakan perpaduan pemodelan linier untuk menangani tugas klasifikasi dalam memecahkan masalah non-linier. Teknik ini berusaha untuk menemukan fungsi pemisah yang optimal yang bisa memisahkan dua kelompok data dari dua kelas yang berbeda.

e. Algoritma Logistic Regression (LR)

Logistic Regression adalah variasi regresi yang digunakan ketika variabel dependen bersifat biner (Yu et al.,2010). Tujuan dari model ini adalah untuk mendapatkan persamaan regresi yang dapat memprediksi dua atau lebih kelompok objek dapat ditempatka apakah pinjaman harus diklasifikasikan sebagai pinjaman yang baik atau pinjaman yang buruk.

f. Algoritma Gradient Boost (GB)

Gradient Boosting merupakan teknik dalam *machine learning* untuk masalah regresi dan klasifikasi yang menghasilkan model prediksi. Pembangunan model dilakukan dengan menggunakan metode *boosting,* yaitu dengan membuat model baru untuk memprediksi *error/residual* dari model sebelumnya. Model baru ditambahkan hingga tidak ada lagi perbaikan pada *error* yang dapat dilakukan. Algoritme ini dinamakan *gradient boosting* karena menggunakan *gradient descent* untuk memperkecil *error* saat membuat model baru.

III. HASIL DAN PEMBAHASAN

A. Dataset

Data yang digunakan adalah data sekunder German Credit Data yang tersedia secara *public* di website kaggle

"https://www.kaggle.com/uciml/germancredit#german_credit_data.csv"/. Total data berisi 1000 entri dengan 10 atribut sebagai berikut :

Tabel 2. Atribut Data

No	Atribut	Keterangan data
1	Age	data numerik (umur)
2	Sex	data kategorikal : - Male - Female
3	Job	data numerik : 0 – unskilled non resident 1 – unskilled-resident 2 – skilled 3 – highly skilled
4	Housing	data kategorikal : - own - rent - free
5	Saving account	data kategorikal : - little - moderate - quite rich - rich
6	Credit amount	data numerik (dalam mata uang German)
7	Checking account	data kategorikal : - little - moderate - quite rich - rich
8	Duration	data numerik (dalam bulan)
9	Purpose	data kategorikal : - car - furniture - radio/tv - domestic appliances - repairs - education - business - vacation/others
10	Risk	data kategorikal : - good - bad

B. Pra Proses Data

Langkah pertama yang dilakukan adalah memeriksa nilai data yang kosong (missing values). Dari gambar 1 dapat diketahui bahwa tidak ada *missing values* pada dataset.

Gambar 1. Hasil running cek tipe data

```
Age                1000 non-null int64
Sex                1000 non-null object
Job                1000 non-null int64
Housing            1000 non-null object
Saving accounts    817 non-null object
Checking account   606 non-null object
Credit amount      1000 non-null int64
Duration           1000 non-null int64
Purpose            1000 non-null object
Risk               1000 non-null object
```

Langkah selanjutnya mentransformasikan atau mengubah nilai data kategorikal menjadi numerik

Tabel 3. Hasil transformasi data

No	Atribut	Kategori	Nilai
1	Sex	- Male - Female	0 1
2	Housing	- own - rent - free	1 2 0
3	Purpose	- car - furniture - radio/tv - domestic appliances - repairs - education - business - vacation/others	1 4 5 2 6 3 7 8
4	Saving account	- little - moderate - quite rich - rich	0 1 3 2
5	Checking account	- little - moderate - quite rich - rich	0 1 3 2
6	Risk	- good - bad	1 0

Setelah tranformasi data menjadi numerik, atur variable x dan y untuk prediksi lalu *split dataset* menjadi data train dan data test.

C. Pengujian

Pengujian dilakukan pada google colab menggunakan bahasa pemrograman python. Perbandingan 4 algoritma klasifikasi dilakukan yaitu Random Forest Classifier (RFC), Support Vector Classifier (SVC), Logistic Regression (LR) dan Gradient Boosting Classifier (GBC) dengan menggunakan library sklearn.

Hasilnya berupa hasil prediksi resiko kredit setiap model yaitu dengan akurasi, *confusion matrix* dan *classification report* (*precision, recall, f1-measure & support*) dan kurva ROC beserta nilai AUC.

D. Hasil Penelitian

Perbandingan data yang digunakan 8:2 yaitu 800 entri data train dan 200 entri data test dengan atribut risk 'good' (1) 141 *value* dan 'bad' (0) 59 *value*, didapatkan hasil *confusion matrix* untuk perhitungan *accuracy*, *precision*, *recall* dan *f-measure* sebagai berikut:

Gambar 2. Confusion Matrix dan hasil klasifikasi RFC

```
Confusion Matrix: [ 27  24  28 121]
Classification report:
              precision    recall  f1-score

           0       0.49      0.53      0.51
           1       0.83      0.81      0.82
```

Gambar 3. Confusion Matrix dan hasil klasifikasi SVC

```
Confusion Matrix: [[ 17  11]
 [ 38 134]]
Classification report:
              precision    recall  f1-score

           0       0.31      0.61      0.41
           1       0.92      0.78      0.85
```

Gambar 4. Confusion Matrix dan hasil klasifikasi LR

```
Confusion Matrix: [[ 21  11]
 [ 34 134]]
Classification report:
              precision    recall  f1-score

           0       0.38      0.66      0.48
           1       0.92      0.80      0.86
```

Gambar 5. Confusion Matrix dan hasil klasifikasi GBC

```
Confusion Matrix: [[ 22  15]
 [ 33 130]]
Classification report:
              precision    recall  f1-score

           0       0.40      0.59      0.48
           1       0.90      0.80      0.84
```

Keterangan TN,FP,FN TP :
False Positive : kemungkinan melunasi pinjaman.
False Negative : kemungkinan tidak melunasi pinjaman.
True Negative : tidak akan melunasi pinjaman.
True Positif : pasti melunaskan pinjaman.

Setelah mendapat hasil dari confusion matrix maka dapat dihitung hasil akurasi setiap algoritma dan juga di berapa lama waktu running program, sebagai berikut:

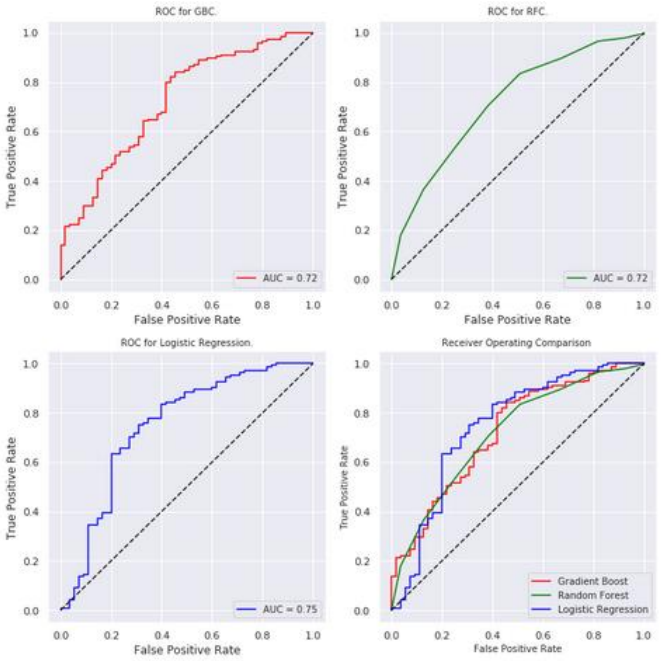
Gambar 6. Hasil akurasi dan kompleksitas waktu

Models	Akurasi	Waktu
Logistic Regression	0.775	0.016103
Gradient Boost Classifier	0.760	0.099944
Support Vector Classifier	0.755	0.038585
Random Forest Classifier	0.740	0.043964

Hasilnya menunjukan dengan Logistic Regression memiliki nilai akurasi terbesar yaitu 0,775 dan kompleksitas waktu lebih cepat dibanding yang lainnya. Di urutan berikutnya, Gradient Boost Classifier dengan nilai akuras 0,760 tetapi kompleksitas waktu tidak sebanding dengan 3 algoritma lainnya dan dianggap kurang efektif.

Selain melihat dari besarnya akurasi dan kompleksitas waktu, nilai AUC untuk menentukan metode mana yang lebih baik digunakan. Nilai AUC adalah luas daerah dibawah kurva ROC. Semakin besar nilai AUC maka semakin baik.

Gambar 7. Kurva ROC dan nilai AUC



Support Vector Classifier tidak bisa dibuat kurva ROC karena kurva ROC dan AUC tidak sensitive terhadap probabilitas prediksi SVC atau dengan kata lain kurvanya akan sama meskipun probabilitas yang diprediksi berkisar 0,9 hingga 1 (bukan dari 0-1). Nilai AUC hanya memperhatikan model klasifikasi untuk membedakan 2 kelas.

Dari kurva diatas dapat diketahui nilai AUC pada Logistic Linear memiliki nilai terbesar dibanding algoritma lainnya dengan nilai 0,75 (klasifikasi cukup).

IV. KESIMPULAN

Penelitian ini melakukan perbandingan 4 algoritma untuk prediksi resiko kredit agar tidak terjadinya resiko gagal bayar yang dilakukan oleh peminjam. Untuk mengukur kinerja algoritma dilakukan evaluasi confusion matrix, akurasi, kompleksitas waktu dan kurva ROC.

Akurasi pada Logistic Regresion (LR) memiliki hasil terbesar dengan nilai 0.775, nilai AUC sebesar 0.75 dan juga kompleksitas waktu paling sedikit (running program tercepat) . Maka dapat disimpulkan dari 4 algoritma klasifikasi untuk prediksi resiko kredit German credit data, Logistic Linear paling cocok untuk digunakan. Kedepannya dibutuhkan data yang lebih banyak karena banyaknya entri dataset menjadi salah satu faktor yang berpengaruh untuk hasil akurasi yang lebih baik.

REFERENSI

[1] D. Kurniawan and D. C. Supriyanto, "Optimasi Algoritma Support Vector Machine (Svm) Menggunakan Adaboost Untuk Penilaian Risiko Kredit," *J. Teknol. Inf.*, vol. 9, no. 1, pp. 1414–9999, 2013.

[2] A. Keramati and N. Yousefi, "A Proposed Classification of Data Mining Techniques in Credit Scoring," *Techniques*, no. September, pp. 416–424, 2011.

[3] S. N. Edusaintek, B. Bawono, R. Wasono, and U. M. Semarang, "PERBANDINGAN METODE RANDOM FOREST DAN NAÏVE BAYES," pp. 343–348, 2019.

[4] C. Algoritma, E. Praja, W. Mandala, and D. E. Putri, "PREDIKSI JUMLAH PEMBERIAN KREDIT KEPADA NASABAH DI BANK PERKREDITAN RAKYAT DENGAN ALGORITMA C 45," vol. 5, no. 1, pp. 70–80, 2018.

[5] I. Menarianti, "Klasifikasi data mining dalam menentukan pemberian kredit bagi nasabah koperasi," *J. Ilm. Teknosains*, vol. 1, no. 1, pp. 1–10, 2015.

[6] D. I. Komputer, F. Matematika, D. A. N. Ilmu, and P. Alam, "Klasifikasi naive bayes pada data tidak seimbang untuk kasus prediksi resiko kredit debitur kartu kredit dewi sri rahayu," 2014.