Perbandingan Metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit

Sri Wahyuningsih[1], Dyah Retno Utari[2]
[1)]AMIK Pakarti Luhur, Banten, Indonesia
[2)] Universitas Budi Luhur, Jakarta, Indonesia
E-mail : [1)]wiwiewahyuningsih@gmail.com, [2)]dyahretno.utari@budiluhur.ac.id

Abstrak

Kendala yang ditemui pada usaha perkreditan adalah kurang akuratnya analisis penilaian debitur terhadap kemampuan dalam melunasi pinjaman kredit, hal tersebut umumnya menyebabkan kredit yang bermasalah. Data Mining dapat digunakan dalam memprediksi kelayakan pemberian kredit untuk calon debitur. Berdasarkan hal tersebut, penelitian ini telah membandingkan metode klasifikasi data mining untuk menganalisis prediksi kelayakan pemberian kredit dengan metode K-NN, Naïve Bayes *dan* Decision Tree. *Data-data calon debitur yang telah melalui tahapan data mining akan diproses menggunakan metode klasifikasi data mining yaitu* K-NN, Naïve Bayes *dan* Decision Tree. *Data akan diuji dengan menggunakan* k-folds cross validation *(k=10). Dari hasil perbandingan tersebut didapat hasil akurasi metode* Decision Tree *(J-48) yang lebih unggul dibandingkan dengan metode K-NN dan* Naïve Bayes. *Hasil yang didapat dari perbandingan ketiga algoritma tersebut adalah, algoritma* Decision Tree *(J-48) dengan akurasi sebesar 92,21%, algoritma* K-Nearest Neighbor *memiliki tingkat akurasi sebesar 81,82% dan algoritma* Naïve Bayes *memiliki tingkat akurasi sebesar 81,83%.*

Kata kunci: *Kelayakan Pemberian Kredit, K-NN, Naïve Bayes, Decision Tree, K-Folds Validation*

1. Pendahuluan

UU Perbankan No. 10 tahun 1998 menerangkan bahwa bank adalah badan usaha yang kegiatannya menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkannya kepada masyarakat dalam bentuk kredit dan atau bentuk – bentuk lainnya dalam rangka meningkatkan taraf hidup rakyat banyak. Berdasarkan UU tersebut, segala bentuk kredit yang dilakukan harus berdasarkan pada persetujuan pinjam-meminjam, dan akan ada suatu analisa yang dilakukan untuk menentukan sebuah pengambilan keputusan.

Kurang tepatnya penilaian awal sebelum menjadi nasabah menjadi permasalahan kredit macet, selain dari BI *checking* sebagai tahap awal seleksi untuk setiap nasabah yang mengajukan kredit. Belum optimalnya pengambilan keputusan dalam hal prediksi kelayakan pemberian kredit kepada nasabah. Permasalahan penelitian dapat dirumuskan sebagai berikut:

1. Apakah metode klasifikasi data mining dapat digunakan untuk memprediksi kelayakan pemberian kredit?
2. Di antara metode K-NN, Naïve Bayes dan Decision Tree, metode klasifikasi apa yang memiliki nilai akurasi tertinggi dalam memprediksi kelayakan pemberian kredit kepada nasabah.

2. Tinjauan Studi

2. Landasan Teori

Data mining atau penambangan data adalah perangkat lunak yang digunakan untuk menemukan pola tersembunyi, tren, maupun aturan-aturan yang terdapat dalam basis berukuran besar dan menghasilkan aturan-aturan yang digunakan untuk memperkirakan perilaku di masa mendatang [1].

2.1. Klasifikasi

Klasifikasi dalam data mining merupakan metode pembelajaran data untuk memprediksi nilai dari sekelompok atribut. Algoritma klasifikasi akan menghasilkan sekumpulan aturan yang disebut rule yang akan digunakan sebagai indikator untuk dapat memprediksi kelas dari data yang ingin diprediksi [2].

2.2. K-Nearest Neighbor (K-NN)

Algoritma K-Nearest Neighbor (K-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan obyek tersebut [3]. Algoritma K-NN menggunakan algoritma supervised.

2.3. Naïve Bayes

Naïve Bayes merupakan machine learning yang menggunakan perhitungan probabilitas yang menggunakan konsep pendekatan Bayesian. Penggunaan teorema Bayes pada algoritma Naïve Bayes yaitu dengan mengkombinasikan prior probability dan probabilitas bersyarat dalam sebuah rumus yang bisa digunakan untuk menghitung probabilitas tiap klasifikasi yang mungkin [4]. Rumus Naïve Bayes adalah:

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

2.4. Decision Tree

Pohon keputusan merupakan pendekatan "divide and conquer" dalam mempelajari masalah dari sekumpulan data independen yang digambarkan dalam bagan pohon [5]. Berikut persamaan data dalam tuple D.

$$Info(D) = \sum_{i=1}^{n} -p_i \, log_2(p_i)$$

p_i merupakan probabilitas tuple dalam D yang menjadi kelas C_i dengan asumsi |C_(i,D) |/|D|. Info(D) atau disebut juga entropy dari D merupakan rata rata informasi yang diperlukan untuk identifikasi tuple dalam D.

2.5. *K-Folds Cross Validation*

Dalam penelitian ini, metode yang digunakan untuk menguji pola klasifikasi adalah dengan metode k-fold cross validation. Dalam k-fold cross validation data dibagi menjadi k bagian, D1,D2…Dk, dan masing-masing D memiliki jumlah data yang sama. Menghitung nilai akurasinya dapat dilakukan dengan menggunakan persamaan:

$$\text{Akurasi} = \frac{Jumlah\ klasifikasi\ benar}{Jumlah\ data\ uji}\ X\ 100\ \%$$

2.6. Pengujian

Untuk menguji model, pada penelitian ini, digunakan metode Confusion Matrix.

1. Confusion matrix

Metode ini menggunakan tabel matriks seperti pada Tabel 1, jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negative [4].

Tabel 1. Model Confusion Matrix

Correct classification	*Clasified as*	
	+	-
+	*True Positives*	*False Negatives*
-	*False Positives*	*True Negatives*

Untuk menghitung digunakan persamaan di bawah ini [6].

Sensitivity = TP/P
Specificity = TN/P
Precision = TP/(TP+FP)
Accuracy = *Sensitivity* P/(P+N) + *Specificity* N/(P+N)

2.7. Penelitian terdahulu

Model Penelitian Emerensye S.Y. Pandie [7]. Hasil pengujian menunjukan bahwa presentase tingkat error data pada angka kurang dari 3,7% dan mencapai kestabilan data pada nilai k=3 sampai k=11 sehingga ditentukan nilai k dengan tingkat error terendah yaitu pada k=7 sebagai nilai k terbaik K-NN untuk prediksi kategori Kredit.

Hasil penelitian lain adalah beberapa algoritma klasifikasi diujicoba terhadap data dengan hasil terbaik pada model klasifikasi dengan algoritma C4.5[9]. Hasil penelitian ini adalah penerapan algoritma C4.5 akan membantu pihak Koperasi dalam menentukan anggota kredit yang akan disetujui pengajuan kreditnya dan menentukan jumlah kredit yang akan dicairkan.

Penelitian lainnya oleh Jayanti dan Noeryanti [10], dari penelitian ini diketahui bahwa ketepatan prediksi resiko kredit dengan metode K-NN adalah sebesar 84,33% pada nilai k=7. Sementara penelitian lainnya oleh Siti Masripah [11], menunjukkan bahwa tingkat akurasi lebih baik saat menggunakan algoritma C4.5 yaitu dengan nilai akurasi 88,90% sedangkan dengan Naïve Bayes 80,00%.

3. Metode Penelitian

Metode penelitian adalah suatu cara atau prosedur untuk mencari, memperoleh, mengumpulkan dan mencatat data yang digunakan dalam menyusun laporan penelitian. Dan akan menggunakan dataset yang dijadikan data training sebanyak 147. Selanjutnya 77 data di luar dari dataset digunakan untuk data *testing*. Data sebanyak 147 records dan terdiri dari 12 atribut, akan dilakukan beberapa penyeleksian untuk menghasilkan data yang dibutuhkan.

3.1. Sampling/Metode Pemilihan Sampel

Penarikan sampel merupakan proses pilihan sejumlah elemen dari populasi. *Probability sampling* adalah teknik pengambilan sampel yang memberikan peluang yang sama bagi setiap unsur atau anggota populasi untuk dipilih menjadi sampel [12]. Untuk data analisis prediksi kredit, didapat data dari Bank ABC dalam kurun waktu 2016, yang terdiri dari 11 atribut, yang terdiri dari 10 atribut adalah *predictor* dan 1 atribut adalah kelas target.

3.2. Metode Pengumpulan Data

Dalam penelitian ini untuk mendapatkan data yang diharapkan maka peneliti mencari, mempelajari, serta mendalami berbagai literatur baik jurnal, buku, ataupun referensi-referensi lainnya yang berhubungan dengan topik penelitian ini.

3.3. Teknik Analisis, Desain, dan Pengujian

Data yang didapat akan dibagi menjadi dua set yaitu sebagai data latih (sekaligus data uji) dan data evaluasi. Hasil dari masing-masing metode dengan data latih akan dibandingkan hasil pengujiannya dengan menggunakan *k-fold cross validation* dengan k=10 untuk mendapatkan hasil berupa nilai akurasi, *precision*, *recall*, *ROC curve*. Algoritma terbaik selanjutnya diimplementasikan menjadi prototipe sistem informasi.

Teknik pengujian terhadap metode yang akan dilakukan menggunakan *k-folds cross validation* dengan k=10. Metode ini membagi data latih secara acak menjadi 10 bagian dengan jumlah yang hampir sama pada masing-masing kelompok. Hasil pengujian akan didapatkan dengan menghitung rata-rata nilai-nilai statistik pengujian pada keseluruhan perulangan.

4. Hasil dan Pembahasan

4.1. Analisis Data

Analisis dilakukan terhadap data dengan sebelas atribut seperti terlihat pada Tabel 2.

Tabel 2. Daftar Atribut Sampel Data

No	Atribut	Keterangan	Status
1	*Gender*/Jenis kelamin	*Female* *Male*	*Predictor*
2	Pendidikan Terakhir	SLTA s/d S3	*Predictor*
3	*Age*	35 s/d 55	*Predictor*
4	Status Perkawinan	Belum Menikah Menikah Janda/Duda	*Predictor*
5	Jumlah tanggungan dalam keluarga	0 s/d 4	*Predictor*
6	Status Tempat Tinggal	Sewa Milik Sendiri Milik Keluarga Rumah Dinas	*Predictor*
7	Pekerjaan	Pegawai Negeri/PNS Pegawai Swasta TNI/Polri Guru Dosen Lainnya	*Predictor*
8	*Income/Year*	65 s/d 11.250	*Predictor*
9	Jumlah Pinjaman yang Diinginkan	50 s/d 5.500	*Predictor*
10	Tujuan Peminjaman	*Personal Loan* *Housing Loan* *Car Loan*	*Predictor*
11	Hasil	*Approval* / Disetujui *Reject* / Ditolak	*Target Class (Label)*

Untuk mengetahui model terbaik, maka hasil perbandingan ukuran evaluasi model klasifikasi akan diperlihatkan dalam nilai-nilai akurasi sebagai berikut.

a. Metode *Decision Tree*
Nilai akurasi dari *confusion matrix* adalah sebagai berikut:

$$= \frac{(66+74)}{(66+5+2+74)}$$
$$\textbf{akurasi} = \mathbf{95,24}\%$$

b. Metode *KNN*
Nilai akurasi dari *confusion matrix* adalah sebagai berikut:

$$= \frac{(57+65)}{(57+14+11+65)}$$
$$\textbf{akurasi} = \mathbf{82,99}\%$$

c. Metode *Naïve Bayes*
Nilai akurasi dari *confusion matrix* adalah sebagai berikut:

$$= \frac{(68+49)}{(68+3+27+49)}$$
$$\textbf{akurasi} = \mathbf{79,59}\%$$

5.1. Evaluasi dan Validasi

a. Confusion Matrix dengan metode *Decision Tree*

Pengujian dilakukan dengan *confusion matrix* yang terdiri dari akurasi *(accuracy*) dilakukan pada 77 *data testing* yang diolah dengan menggunakan ketiga metode. Model dengan algoritma *Decision Tree, K-NN* dan *Naïve Bayes* untuk prediksi kelayakan pemberian kredit kepada nasabah yang diuji tingkat akurasinya menghasilkan perbandingan nilai akurasi (*accuracy*). Dengan dataset nasabah sebagai data uji. Algoritma *Decision Tree* mendapatkan akurasi yang paling tinggi yaitu sebesar 93,72%, Algoritma *K-NN* mendapatkan akurasi sebesar 82,41%. Sedangkan algoritma *Naïve Bayes* mendapatkan akurasi sebesar 80,71%. Seperti yang ditunjukkan pada tabel 10.

Tabel 6. Perbandingan Tingkat Akurasi *Decision Tree, K-NN* dan *Naïve Bayes*

Metode Data Mining	Nilai Akurasi
Decision Tree	**92,21%**
K-Nearest Neighbor	81,82%
Naïve Bayes	81,83%

5.2. Penerapan Algoritma Terpilih

Dari hasil evaluasi dan validasi di atas dapat diketahui bahwa metode *Decision Tree* memiliki tingkat akurasi dan perfomansi yang baik, sehingga rule yang dihasilkan oleh metode *Decision Tree* dapat dijadikan sebagai *rule* untuk pembuatan *prototype* yang dapat memudahkan dalam memprediksi kelayakan pemberian kredit kepada calon nasabah. Prototipe yang diusulkan memiliki rancangan seperti terlihat pada Gambar 4. Proses validasi pada prototipe yang dibangun dengan algoritma terbaik yaitu *Decision Tree* menunjukkan nilai akurasi yang sama dengan pengujian model, yaitu sebesar 92,21%.

Gambar 5. Implementasi Sistem Informasi

6. Simpulan dan Saran

Metode *Decision Tree* memiliki tingkat akurasi yang baik yaitu sebesar 92,21% untuk prediksi kelayakan pemberian kredit kepada nasabah, metode *K-Nearest Neighbor* sebesar 81,82% dan metode *Naïve Bayes* memiliki akurasi sebesar 81,83%. Meskipun algoritma *Decision Tree* memiliki nilai akurasi yang tinggi, namun dapat dilakukan pengembangan selanjutnya, hal-hal berikut bisa ditambahkan untuk meningkatkan akurasi dan performa yaitu:

1. Mengkombinasikan lebih banyak metode dalam Analisa data dan penyelesaian masalah, sehingga didapat sebuah sistem yang lebih efektif dan efisien dalam pengolahan ataupun penyajian informasi.
2. Pengelolaan waktu penelitian agar dapat lebih dimaksimalkan, mengingat pendeknya waktu yang tersedia.
3. Peran dari responden sangatlah penting dalam mendukung penelitian ini. Terutama respoden yang berkaitan langsung dengan objek penelitian.
4. Penelitian ini dapat dikembangkan dengan algoritma klasifikasi yang lain yang terdapat dalam data mining, seperti algoritma *Neural Network*, *K-Means* atau *SVM (Support Vector Machine*).

Daftar Pustaka

[1] A. Kadir, *Pengenalan Sistem Informasi Edisi Revisi*, no. Penerbit Andi. Yogyakarta: Andi, 2014.
[2] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making (Google eBook)*, no. 2004. 2011.
[3] F. Gorunescu, *Data Mining Concepts, Models and Techniques*. Springer-Verlag, 2011.
[4] M. Bramer, *Principles of Data Mining*. Springer-Verlag London, 2013.
[5] I. H. Witten, E. Frank, and M. a. Hall, *Data Mining Practical Machine Learning Tools and Techniques Third Edition*, vol. 277, no. Tentang Data Mining. 2011.
[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
[7] E. S. Y. Pandie, "Implementasi Algoritma Data Mining K-NN dalam pengambilan keputusan Pengajuan Kredit," 2012.
[8] H. Marcos, I. Hidayah, J. Teknik, E. Dan, T. Informatika, and U. G. Mada, "IMPLEMENTASI DATA MINING UNTUK KLASIFIKASI NASABAH KREDIT BANK ' X ' MENGGUNAKAN CLASSIFICATION RULE," pp. 1–7, 2014.
[9] S. A. Lusinia, S. Kom, M. Kom, and F. I. Komputer, "Algoritma C4.5 dalam menganalisa kelayakan kredit(studi kasus di koperasi pegawai Republik Indonesia(KP-RI))Lengayang Pesisir Selatan, Painan, Sumatera Barat," vol. 1, no. 2, pp. 6–10, 2014.
[10] J. R. Dwi and Noeryanti, "Aplikasi Metode K-Nearest Neighbor dan Analisis diskriminan untuk analisa resiko kredit pada koperasi simpan pinjam di Kopinkra Sumber Rejeki," pp. 275–284, 2014.
[11] S. Masripah, "Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit," vol. 3, no. 1, pp. 187–193, 2016.
[12] S. Guritno, Sudaryono, and U. Rahardja, *Theory and Application of IT Research-Metodologi dan Penelitian Teknologi Informasi*. Yogyakarta: Andi, 2011.