Data Mining Klasifikasi Untuk Memprediksi Status Keberlanjutan Polis Asuransi Kesehatan Dengan Algoritme Naïve Bayes

Jovansgha Avegad dan Arief Wibowo
Universitas Budi Luhur
Jl. Ciledug Raya, Petukangan Utara, Jakarta Selatan 12260
jovansghaavegad@gmail.com, arief.wibowo@budiluhur.ac.id

ABSTRAK

Di saat semakin kompetitifnya persaingan bisnis asuransi komersial kategori non-jaminan kesehatan nasional (JKN), membuat perusahaan asuransi umum dituntut memiliki inovasi dalam meningkatkan jumlah pelanggan atau nasabahnya. Di sisi lain, keputusan nasabah asuransi kesehatan terhadap keberlanjutan polis yang telah dimiliki apakah akan diperpanjang atau tidak diperpanjang, tidaklah mudah untuk diprediksi. Untuk dapat memprediksi keputusan nasabah asuransi kesehatan dalam status keberlanjutan polis maka diperlukan suatu metode analisis data pelanggan asuransi yang telah terdaftar. Penelitian ini menguraikan pemanfaatan algoritme Naïve Bayes sebagai salah satu algoritme terbaik dalam data mining klasifikasi untuk memprediksi keputusan nasabah asuransi terhadap polis yang dimiliki. Berdasarkan hasil pemodelan yang dilakukan, diketahui bahwa algoritme klasifikasi Naïve Bayes mampu memprediksi status keberlanjutan polis asuransi dengan akurasi mencapai 88,00%, presisi 89,19% dan recall 100%. Dengan hasil yang relatif baik ini maka dapat dilakukan upaya-upaya peningkatan pendapatan perusahaan asuransi kesehatan misalnya dengan menawarkan program promosi pembaharuan polis asuransi pada nasabah-nasabah yang diprediksi akan memperpanjang maupun yang tidak akan memperpanjang polis asuransi yang dimilikinya.

Kata Kunci : *Data Mining, Klasifikasi, Naïve Bayes, Prediksi Pembaharuan Polis Asuransi Kesehatan*

PENDAHULUAN

PT XYZ merupakan perusahaan yang bergerak dalam bidang asuransi dan memiliki nasabah yang cukup banyak. Data nasabah dari PT XYZ belum diolah sedemikian rupa untuk dimanfaatkan dalam prediksi pembaharuan polis asuransi.

Dengan proses Data *Mining* dari data nasabah, dapat ditemukan pola-pola ataupun hubungan keterkaitan tertentu antara data untuk menjadi informasi yang berharga [1]. Adapun tahapan mendapatkan informasi ini dapat diperoleh dalam tahap *Knowledge Discovery in Database* (KDD).

Penelitian ini menguraikan bagaimana metode data *mining* algoritme Naïve Bayes *atau yang* disebut juga Bayesian *Classification* mampu digunakan untuk mengklasifikasi probabilitas status keberlanjutan polis asuransi kesehatan.

Naïve Bayes *Classifier* didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa *decision tree* dan *neural network*. Selain itu, Naïve Bayes *Classifier* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database yang besar [2]. Adapun persamaan atau formula yang digunakan dalam algoritme menurut teori Bayes tersebut adalah:

$$(H \mid X)=(P(X \mid H)P(H))/P(X) \quad (1)$$

Dimana :

X = Data dengan class yang belum diketahui
H = Hipotesis data X merupakan suatu class spesifik.
P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (Posteriori Probability)
P(H) = Probabilitas hipotesis H (Prior Probability)
P(X|H) = Probabilitas X berdasarkan hipotesis H
P(X) = Probabilitas dari X

Dalam proses evaluasi dari model yang dihasilkan, digunakan alat analisis *confusion matrix*, untuk mendapatkan informasi akurasi, presisi dan *recall* dari hasil pemodelan [3].

a. Akurasi
   Akurasi dalam klasifikasi adalah persentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi.
b. Presisi
   Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut.
c. Recall
   Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia.

Hasil pengujian didapatkan dengan cara *holdout validation*, yaitu dengan membagi sebagian dataset menjadi data latih dan data uji.

METODE PENELITIAN

Tahapan metode penelitian yang digunakan yang direferensi berdasarkan tahapan dalam KDD adalah sebagai berikut:

1) *Data Selection*
   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data operasional.
2) *Pre-Processing/Cleaning*
   Proses *cleaning* antara lain membuang duplikasi data, memeriksa data yang tidak konsisten dan memperbaiki kesalahan pada data. Pada proses ini dilakukan juga proses enrichment, yaitu proses memperkaya data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD.
3) *Transformation*
   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining.
4) Data *Mining*
   Data *Mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.
5) *Interpretation/Evaluation*
   Pola informasi yang dihasilkan dari proses data *mining* diterjemahkan menjadi bentuk yang lebih mudah dimengerti oleh pihak yang berke-pentingan.

HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan seluruh tahapan yang telah dilakukan, mulai dari *data selection* hingga *interpretation* atau *evaluation*

Data Selection

Dataset yang diterima terdiri dari beberapa atribut yaitu:

1) No polis
2) Nama
3) Jenis kelamin
4) Usia
5) Pekerjaan
6) Tanggungan
7) Mulai asuransi
8) Plan
9) Premi tahunan
10) Periode asuransi
11) Riwayat penyakit
12) Rasio klaim
13) Perpanjangan

Dari atribut-atribut data yang dimiliki maka dilakukan pemilihan atribut sebagai acuan klasifikasi. Atribut yang dipilih sebagai berikut:

1) Jenis kelamin
2) Usia
3) Pekerjaan
4) Tanggungan
5) Plan
6) Premi tahunan
7) Riwayat penyakit
8) Rasio klaim
9) perpanjang

Pre-Processing/Cleaning

Tahap ini membersihkan data dari duplikasi dan data yang tidak konsisten, atau dikatakan tidak bersih, misalnya menghapus data yang tidak lengkap.

Transformation

Dilakukan transformasi pada beberapa data agar dapat diproses dengan metode

yang digunakan dalam data *mining*, transformasi data sebagai berikut:

Tabel 1. *Transformasi data usia*

Usia	Transformasi
1 – 28	KategoriUsia1
29 - 38	KategoriUsia2
39 - 48	KategoriUsia3
49 - 99	KategoriUsia4

Pembagian range usia ini diperoleh dari klasifikasi yang diberlakukan pada lokasi obyek penelitian. Transformasi lainnya yang dilakukan adalah:

Tabel 2. *Transformasi data tanggungan*

Tanggungan	Transformasi
0	Tanpa Tanggungan
1 - Seterusnya	Ada Tanggungan

Dengan jumlah tanggungan yang berisi nilai 0 dan seterusnya, maka dapat dilakukan kategorisasi tanggungan menjadi nilai 'Ada tanggungan' dan 'Tanpa tanggungan'.

Premi tahunan juga ditransformasi berdasarkan median terhadap setiap plan yang ada. Transformasi juga dilakukan pada nilai premi sebagai berikut:

Tabel 3. *Kelompok transformasi premi*

Plan	Range (IDR)	
1	>822.600	<=822.600
2	>1.936.000	<=1.936.000
3	>3.335.400	<=3.335.400
4	>4.728.000	<=4.728.000
5	>4.728.000	<=4.728.000

Data *Mining*

Data mining adalah proses pencarian informasi didalam data dengan suatu metode yaitu Naïve Bayes.

Pemodelan ini menggunakan dataset faktual dari perusahaan asuransi kesehatan di DKI Jakarta pada tahun 2018 sejumlah 225 record yang dibagi menjadi 2/3 data latih dan 1/3 data uji sebagaimana metode validasi *holdout validation* [4].

Berikut ilustrasi penerapan analisis satu buah record dari dataset menggunakan algoritme Naïve Bayes:

a. Jenis Kelamin : L
b. Usia : dewasa awal
c. Pekerjaan : Pegawai Swasta
d. Tanggungan : ada
e. Plan : D
f. Premi tahunan : <=1936000
g. Riwayat Penyakit : rendah
h. Rasio Klaim : sedang
i. Status : renewal

Berdasarkan persamaan dari penerapan Algoritme Naïve Bayes maka diperoleh hasil sebagai berikut:

1. Kelas spesifik :
   - P(renewal) = 0.91
   - P(tidak renewal) = 0.09
2. Jenis Kelamin :
   - P(L|renewal) = 0.47
   - P(L|tidak renewal) = 0.38
   - P(L) = 0.47
3. Usia
   - P(dewasa awal|renewal) = 0.26
   - P(dewasa awal|tidak renewal) = 0.46
   - P(dewasa awal) = 0.28
4. Pekerjaan
   - P(Pegawai Swasta|renewal) = 0.58
   - P(Pegawai Swasta|tidak renewal) = 0.46
   - P(Pegawai Swasta) = 0.57
5. Tanggungan
   - P(ada|renewal) = 0.41
   - P(ada|tidak renewal) = 0.15
   - P(ada) = 0.39
6. Plan
   - P(D|renewal) = 0.55
   - P(D|tidak renewal) = 0.54
   - P(D) = 0.55
7. Premi Tahunan
   - P(<=1936000|renewal) = 0.31
   - P(<=1936000|tidak renewal) = 0.08
   - P(<=1936000) = 0.29
8. Riwayat Penyakit
   - P(rendah|renewal) = 0.42
   - P(rendah|tidak renewal) = 0.54
   - P(rendah) = 0.43
9. Rasio Klaim
   - P(sedang|renewal) = 0.35

- P(sedang|tidak renewal) = 0.46
- P(sedang) = 0.36

Probabilitas *renewal*:

$$(renewal|X) = \frac{P(X|renewal)P(renewal)}{P(X)} \quad (2)$$

Dimana X = Data dengan *class* yang belum diketahui.

P(X|renewal) = P(L|renewal)×P(dewasa awal|renewal) × P(Pegawai Swasta|renewal) × P(ada|renewal) × P(D|renewal) × P (≤1936000|renewal) × P(rendah|renewal) × P(sedang|renewal)

P(X|*renewal*) = 0.47×0.26×0.58×0.41× 0.55×0.31×0.42×0.35

P(X|*renewal*) = 0.000728324257

P(X) = P(L)×P(dewasa awal)×P(Pegawai Swasta)×P(ada)×P(D)×P(≤193600 0)×P(rendah)×P(sedang)

P(X) = 0.47×0.28×0.57×0.39×0.55×0.29× 0.43×0.36
P(X) = 0.000223261186

P(*renewal*) = 0.91

$$(renewal|X) = \frac{0.000728324257 \times 0.91}{0.000722415602}$$

$$(renewal|X) = 0.917442913$$

Probabilitas tidak *renewal*:

$$(tidak\ renewal|X) = \frac{P(X|tidak\ renewal)P(tidak\ renewal)}{P(X)} \quad (3)$$

Dimana X adalah data dengan *class* yang belum diketahui.

P(X|tidak renewal) =
P(L | tidak renewal)×P(dewasa awal | tidak renewal)×P(Pegawai Swasta | tidak renewal)×P(ada | tidak renewal)×P(D | tidak renewal)×P(≤1936000 | tidak renewal)×P(rendah | tidak renewal)×P(sedang | tidak renewal)

P(X | tidak renewal)
= 0.38×0.46×0.46×0.15×0.54×0.08× 0.54×0.46

P(X | tidak renewal)
=0.00012942729
P(X) = P(L)×P(dewasa awal)×P(Pegawai Swasta)×P(ada)×P(D)×P(≤193600 0)×P(rendah)×P(sedang)

P(X) = 0.47×0.28×0.57×0.39×0.17×0.29× 0.43×0.36

P(X) = 0.000223261186

P(*tidak renewal*) = 0.09

$$(\text{tidak}\ renewal|X) = \frac{0.00012942729 \times 0.09}{0.000722415602}$$

$$(\text{tidak}\ renewal|X) = 0.0161243141$$

Tahap akhir adalah membandingkan nilai probabilitas *renewal* dengan probabilitas tidak *renewal*:

(*renewal*|X) = 0.917442913
(tidak *renewal*|X) = 0.0161243141
(*renewal*|X) > (tidak *renewal*|X)

Dengan diketahuinya probabilitas *renewal* yang lebih besar, maka hasil analisis data dengan algoritme Naïve Bayes memberikan hasil prediksi bahwa keputusan nasabah terhadap polis asuransi yang dimiliki adalah akan diperbaharui (*renewal).*

Interpretation/Evaluation

Dilakukan proses evaluasi untuk merubah hasil informasi menjadi data yang lebih mudah dipahami, evaluasi dilakukan dengan metode *confusion matrix.*

Hasil pengujian terhadap 1/3 dataset sebagai data uji, mendapatkan nilai akurasi sebesar 88,0%, presisi 89,19% dan *recall* 100% sebagaimana tabel berikut:

Tabel 3. *Hasil Confusion Matrix*

Data Aktual	Hasil Prediksi	
	Renewal	**Tidak Renewal**
Renewal	66 (A)	1 (B)
Tidak Renewal	8 (C)	0 (D)

Perhitungan *Confusion Matrix,* sebagai berikut:

a. Perhitungan Akurasi

$$Akurasi = \frac{A + D}{A + B + C + D}$$

$$Akurasi = \frac{66 + 0}{66 + 8 + 1 + 0}$$

$$Akurasi = 0.88$$

b. Perhitungan Presisi

$$Presisi = \frac{A}{C + A}$$

$$Presisi = \frac{66}{8 + 66}$$

$$Presisi = 0.8919$$

c. Perhitungan *Recall*

$$Recall = \frac{A}{A + D}$$

$$Recall = \frac{66}{66 + 0}$$

$$Recall = 1$$

Dengan pengujian yang dilakukan didapatkan hasil akurasi yang relatif baik yaitu mencapai 88,0% disertai presisi sebesar 89,19% dan Recall sebesar 100%.

PENUTUP

Kesimpulan analisis dari pengujian prediksi perpanjangan polis asuransi dengan algoritme Naïve Bayes, memiliki akurasi 88,0%, presisi 89,19% dan *recall* 100%. Hasil yang relatif baik dan akurat ini memungkinkan bagi PT XYZ dapat melakukan upaya-upaya peningkatan pendapatan misalnya dengan menawarkan program perpanjangan polis asuransi pada nasabah-nasabah tertentu yang diprediksi akan memperpanjang polisnya, atau memberikan penawaran menarik agar nasabah yang diprediksi tidak memperpanjang polisnya menjadi keputusan untuk memperbaharui/memperpanjang polis yang telah dimilikinya.

Saran

Saran untuk mengembangkan penelitian terhadap data nasabah asuransi lebih lanjut:

a) Penelitian berikutnya dapat dikembangkan dengan menggunakan metode yang mungkin lebih baik, seperti C4.5 atau k-NN.
b) Penelitian berikutnya dapat dikembangkan dengan metode validasi yang mungkin lebih baik, seperti K-*fold validation.*
c) Pengumpulan data diharapkan lebih banyak dan bervarasi sehingga diharapkan akan meningkatkan nilai akurasi.

DAFTAR PUSTAKA

[1] . N. Nuraeni, "Penentuan Kelayakan Kredit Dengan Algoritma Naïve Bayes Classifier : Studi Kasus Bank Mayapada Mitra Usaha Cabang PGC," J. Tek. Komput. AMIK BSI, vol. 3, no. 1, pp. 9–15, 2017.

[2] . Y. V. Via, B. Nugroho, and A. Syafrizal, "Sistem Pendukung Keputusan Klasifikasi Tingkat Keganasan Kanker Payudara Dengan Metode Naive Bayes Classifier," SCAN-Jurnal Teknol. Inf. dan Komun., vol. 10, no. 2, pp. 63–68, 2015.

[3] . E. Mayadewi, P., & Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining," Sist. Inf. Indones., no. November, pp. 2–3, 2015.

[4] . T. Praningki and I. Budi, "Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN," Creat. Inf. Technol. J., vol. 4, no. 2, p. 83, 2018.

[5] . Bustami, "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," J. Inform., vol. 8, no. 1, pp. 884–898, 2014