MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

HETEROGENEOUS MULTIPLE CLASSIFIERS MENGUNAKAN C4.5, K-NEAREST NEIGHBOR DAN NAÏVE BAYES UNTUK MENENTUKAN TINGKAT PEMBAHARUAN POLIS ASURANSI JIWA

[1)] Reni Utami
Universitas Dian NusantaraTeknik dan Informatika, Kampus 1 Tanjung Duren
E-Mail: reni.utami@dosen.undira.ac.id

[2)] Irfan Nurdiansyah
Universitas Dian NusantaraTeknik dan Informatika, Kampus 1 Tanjung Duren
E-Mail: irfan.nurdiansyah@dosen.undira.ac

At a time when the insurance business is increasingly competitive, it requires insurance companies to have innovations in increasing the number of customers. With information from existing customer data, insurance companies can make decisions in implementing company strategies, including determining insurance customer decisions on the sustainability of life insurance policies. Data mining can form a pattern or create a trait of business behavior that is useful for decision making. In this research a Heterogeneous Multiple Classifiers prediction model was built using Majority Voting by combining C4.5, K-Nearest Neighbor and Naïve Bayes to determine the renewal rate of life insurance policies. The Heterogeneous Multiple Classifiers model that was built produced an accuracy value of 94.61%, precision value of 95.20%, recall value of 94.60% and an F-Measure value of 94.60%. The performance value generated by the Heterogeneous Multiple Classifiers based prediction model is higher than the performance value of the Single Classifier based prediction model. It is hoped that this method can increase the income of life insurance companies, for example by offering a promotional program for insurance policy renewal to customers who are predicted to extend or not to extend their insurance policies.

Keyword : Heterogeneous Multiple Classifiers, C4.5, K-Nearest Neighbor, Naïve Bayes, Majority Voting, Insurance Policy Renewal

PENDAHULUAN

Komponen Perkembangan teknologi informasi saat ini yang semakin maju dan pesat sangat berarti bagi semua kalangan masyarakat, sesuai dengan tingkat kebutuhan manusia untuk memecahkan permasalahannya. Saat ini teknologi informasi telah menjadi salah satu kebutuhan dalam kehidupan sehari-hari yang digunakan untuk merespon data transaksi dan menyediakan informasi untuk mendukung pengambilan keputusan. Pemanfaatan teknologi informasi terbukti dapat mempermudah kinerja manusia dalam menjalankan suatu pekerjaan. Hal inilah yang menyebabkan teknologi informasi diterapkan dalam beragam bidang yang ada, tidak terkecuali dalam dunia bisnis asuransi. Layanan nasabah merupakan kunci dalam dunia usaha asuransi untuk terus berkembang dan berinovasi dalam memberikan suatu pelayanan terhadap produk yang dijual, maka perusahaan harus dapat membentuk dan menerapkan terobosan baru agar dapat bersaing yakni dengan menggunakan Fasilitas Teknologi Informasi agar dapat memberikan pengaruh positif pada layanan nasabah dan kinerja perusahaan. "Pertumbuhan dapat memungkinkan organisasi untuk memperoleh dari skala keuntungan, untuk meningkatkan posisinya diantara pesaing industrinya, dan memberikan lebih banyak kesempatan untuk pengembangan profesional dan kemajuan kepada karyawan (Mello, 2010)".

Teknologi data mining dapat dimanfaatkan pada data pelanggan asuransi jiwa dan data transaksi yang terjadi pada premi asuransi maupun data klaim. Dari data tersebut, akan diklasifikasi pelanggan-pelanggan yang potensial dalam memperbaharui polis (renewal) atau menghentikan polis dalam pengajuan klaim asuransi yang dimiliki nasabah. Klasifikasi dapat diartikan sebagai sebuah proses menemukan suatu model atau fungsi yang menggambarkan dan membedakan kelas objek data, dengan tujuan untuk menggunakan model yang dihasilkan pola – pola dalam pengambilan keputusan para pelanggan dalam melakukan renewal atau pembaharuan polis asuransi.

Data Mining merupakan teknologi yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari Gudang data mereka. Data Mining meramalkan trend dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. Analisis yang diotomatisasi yang dilakukan oleh data mining melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan. Data Mining mengeksplorasi basis

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

data untuk menemukan pola-pola yang tersembunyi, mencari informasi pemrediksi yang mungkin saja terlupakan oleh para pelaku bisnis karena terletak di luar ekspektasi mereka (Santosa, 2007).

Berdasarkan hasil analisa dan tinjauan studi literatur penelitian menjadi latar belakang penulis dalam melakukan penelitian pada penulisan tesis yang berjudul "Heterogeneous Multiple Classifiers mengunakan C4.5, K-Nearest Neighbor dan Naïve Bayes untuk menentukan tingkat pembaharuan polis asuransi jiwa.

METODE PENGUMPULAN DATA

Metode pengumpulan data yang tepat yaitu dengan mempertimbangkan penggunaannya berdasarkan jenis data dan sumbernya. Data yang objektif dan relevan dengan pokok permasalahan penelitian merupakan indikator keberhasilan suatu penelitian. Pengumpulan data penelitian yang dilakukan dengan menggunakan sebagai berikut.

Observasi merupakan metode pengumpulan data dengan cara mengadakan pengamatan langsung kepada objek penelitian mengenai pendataan nasabah yang terjadi pada perusahaan asuransi jiwa. Serta sumber datanya berasal dari data sekunder yang diperoleh dalam bentuk yang telah menjadi informasi seperti dataset pada database. Pada penelitian ini data sekunder yang didapatkan yaitu data pemegang polis seluruh wilayah Indonesia pada tahun 2015 hingga 2018.

Wawancara merupakan teknik pengumpulan data dengan cara mengadakan tanya jawab atau wawancara langsung kepada bagian terkait secara menyeluruh apa yang dibutuhkan terkait pembuatan penulisan akhir. Serta sumber datanya berasal dari data primer yang diperoleh dari secara langsung ke bagian objek yang diteliti atau berasal dari sumber orangnya.

Studi pustaka mengumpulkan data dengan mempelajari masalah yang berhubungan dengan objek yang diteliti serta bersumber dari buku – buku.

Teknik Analisis/Rancangan dan Pengujian Data/ Sistem/ Prototipe/ Model/ Rencana Strategi

Rencana Strategi yang dilakukan oleh peneliti yaitu berdiskusi langsung kepada unit terkait mengenai cara penentuan atribut yang digunakan pada saat explorasi data nanti. Atribut yang dipilih bedasarkan pakar dari unit klaim dan marketing. Atribut yang digunakan pada saat pengelolaan data dapat dilihat pada Tabel 1.

Tabel 1. Atribut yang Digunakan

Atribut	Range Data / Keterangan
Usia	Usia pemegang polis saat membuka polis baru / Dalam satuan tahun
Jumlah anggota keluarga	Jumlah anggota keluarga yang Ter-*cover* pada polis / 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, dst
Total UP *(Sum insured)*	Total uang pertanggungan berdasarkan jumlah anggota keluarga / Dalam satuan rupiah
Klaim frekuensi	Berdasarkan intensitas pengajuan klaim pada satu polis / 1, 2, 3, dst
Pilihan paket	Jenis pilihan paket yang dipilih oleh pemegang polis / A1, A2, A3, A4, A5, B1, B2, B3, B4, B5
Jenis klaim (Penyebab meninggal)	Penyebab meninggal yang diajukan saat klaim / Sakit, kecelakaan, lainnya, -
Renewal	Status dari pembaharuan polis yang telah *expired* / Y, T

Menurut Han Jiawei (2012) data preprocessing diperlukan dalam proses data mining dikarenakan data yang tidak berkualitas akan menghasilkan kualitas mining yang tidak baik. Data preprocessing, cleaning, dan transformasi merupakan pekerjaan mayoritas dalam aplikasi data mining. *Preprocessing* data diperlukan pada data mining. *Noisy*, yaitu data yang masih mengandung error dan outliers. *Incomplete*, yaitu data yang kekurangan nilai atribut atau hanya mengandung agregat data, artinya adanya atribut yang tidak memiliki nilai. *Incosistency*, yaitu data yang mengandung discrepansi dalam code dan nama atau singkatnya data yang tidak konsisten.

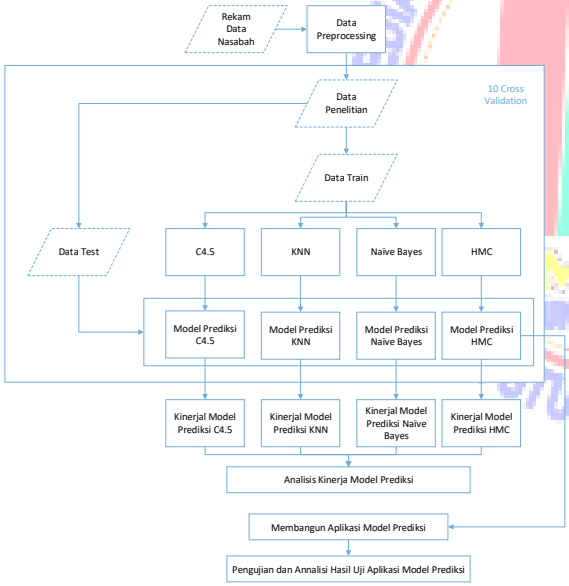
Ada beberapa hal yang dapat dilakukan untuk preprocessing data mining. Data validation adalah Tahapan untuk mengidentifikasi dan menghapus data yang ganjil (noise), data yang tidak konsistensi dan data yang tidak lengkap (missing value). Data yang menjadi acuan adalah no. kartu keluarga dan no. polis sebagai identifikasi dari pemegang polis. Data Integration and Transformation adalah Tahapan untuk meningkatkan akurasi dan efisiensi algoritma. Data yang diolah yaitu jumlah frekuensi pengajuan klaim yang lebih dari sama dengan 1 pengajuan klaim dan yang tidak melakukan pengajuan klaim serta yang menjadi acuan adalah no. polis dan pilihan paket yang dipilih.Integration. menggabungkan beberapa sumber data sehingga dapat saling melengkapi. data perlu digabungkan dengan key yang sesuai. key ini mungkin memiliki nama yang berbeda di sumber data yang berbeda. Data Size Reduction and Discretization adalah Tahapan untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informatif. Perlu dilakukan diskritisasi (memecahkan domain atau daerah perhitungan menjadi beberapa daerah-daerah kecil) karena kolom pada data diatas memiliki range yang cukup luas. Aturan range ini dapat diubah sesuai dengan keinginan peneliti. Berikut adalah contoh beberapa nilai diskritisasi variabel dari pilihan paket.

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

Tabel 2. *Discretization* Variabel Pilihan Paket

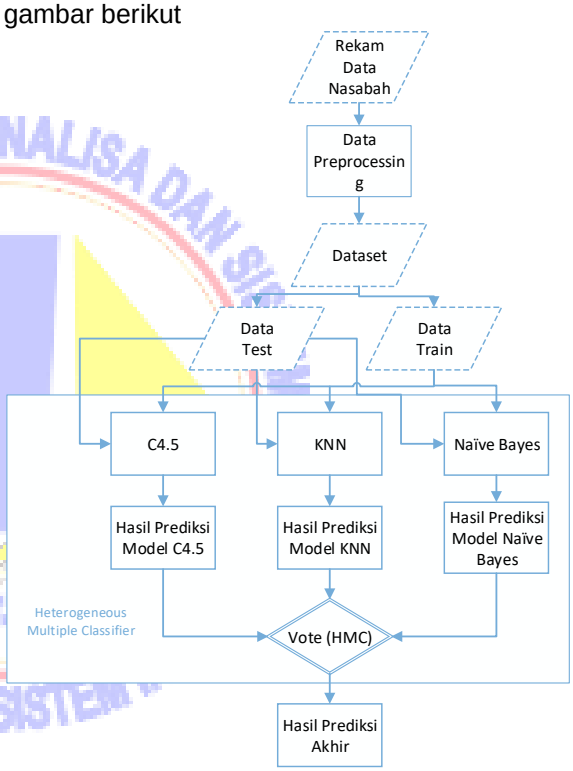
NO.	PILIHAN_PAKET	HUB_KELUARGA	SANTUNAN	PREMI
1	A1	KPLKLGA	1.500.000	50.000
2	A1	ISTRI	1.500.000	50.000
3	A1	ANAK	1.500.000	50.000
4	A2	KPLKLGA	3.250.000	100.000
5	A2	ISTRI	3.250.000	100.000
6	A2	ANAK	3.250.000	100.000
7	A3	KPLKLGA	4.750.000	150.000
8	A3	ISTRI	4.750.000	150.000
9	A3	ANAK	4.750.000	150.000
10	A4	KPLKLGA	6.500.000	200.000
11	A4	ISTRI	6.500.000	200.000
12	A4	ANAK	6.500.000	200.000
13	A5	KPLKLGA	8.000.000	250.000
14	A5	ISTRI	8.000.000	250.000
15	A5	ANAK	8.000.000	250.000
16	B1	KPLKLGA	2.000.000	50.000
17	B1	ISTRI	1.500.000	50.000
18	B1	ANAK	1.000.000	50.000
19	B2	KPLKLGA	4.250.000	100.000
20	B2	ISTRI	3.250.000	100.000
21	B2	ANAK	2.000.000	100.000
22	B3	KPLKLGA	6.500.000	150.000
23	B3	ISTRI	4.750.000	150.000
24	B3	ANAK	3.000.000	150.000
25	B4	KPLKLGA	8.500.000	200.000
26	B4	ISTRI	6.500.000	200.000
27	B4	ANAK	4.000.000	200.000
28	B5	KPLKLGA	10.000.000	250.000
29	B5	ISTRI	9.000.000	250.000
30	B5	ANAK	5.000.000	250.000

Rancangan penelitian Model prediksi Penentuan Pembaharuan Polis Asuransi Berbasis Heterogeneous Multiple Classifiers ini, dapat dilihat pada gambar berikut.



Gambar 1. Rancangan Penelitian

Dari gambar di atas, dapat dijelaskan bahwa data rekam nasabah sebelum digunakan sebagai data penelitian, akan dilakukan Data Preprocessing yang meliputi data validation, integration dan discretization. Selanjutnya data penelitian akan dibagi menjadi 2 bagian yaitu data train dan data test. Data train akan digunakan untuk membangun model prediksi. Data uji digunakan untuk menguji model prediksi yang telah dibangun. Model prediksi yang dibangun terdiri dari empat model, yaitu tiga model prediksi berbasis Single Classifier (model prediksi menggunakan Naïve Bayes, model prediksi menggunakan KNN, model prediksi menggunakan C4.5) dan model prediksi berbasis Heterogeneous Multiple Classifiers (kombinasi Naïve Bayes, KNN dan C4.5 dengan menggunakan metode penggabungan Simple Majority Vote). Model prediksi-model prediksi yang telah dibangun, kemudian diujikan dengan menggunakan data uji, untuk dihitung kinerja model prediksi-model prediksi tersebut. Pembangunan dan pengujian model prediksi-model prediksi tersebut dilakukan dengan 10 fold cross validation. Kinerja model prediksi Naïve Bayes, model prediksi KNN dan model prediksi C4.5 akan dianalisis dengan membandingkan kinerja yang dihasilkan dari model prediksi berbasis Heterogeneous Multiple Classifiers. Rancangan model prediksi penentuan pembaharuan polis asuransi jiwa berbasis Heterogeneous Multiple Classifiers yang akan dibangun dalam penelitian ini, dapat dilihat pada gambar berikut



Gambar 2. Model Prediksi Renewal Polis Asuransi Jiwa

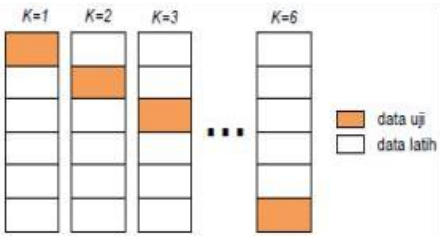
Dari gambar di atas dapat dijelaskan bahwa model prediksi berbasis Heterogeneous Multiple Classifiers yang dibangun pada penelitian ini menggabungkan tiga classifier yaitu Naive Bayes, KNN dan C4.5. Hasil akhir dari prediksi model ini adalah hasil penggabungan dari prediksi ketiga classifier tersebut dengan menggunakan Majority Voting

Pengujian dan Pengukuran

Untuk menguji model pada penelitian ini, digunakan metode K-Fold Cross Validation dan Confusion Matrix. K-Fold Cross Validation merupakan salah satu dari variasi teknik

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

pengujian cross validation. k-fold cross validation dilakukan dengan membagi training set dan test set. Keuntungan menggunakan k-fold cross validation dibandingkan dengan variasi cross validation seperti repeated random sub-sampling validation adalah semua data akan digunakan, baik untuk data uji maupun untuk data latih. Sebagai ilustrasi dari k-fold cross validation dapat dilihat pada Gambar 2.8 (Payam et al, 2009).



Gambar 3. Ilustrasi K-Fold Cross Validation

Dalam ilustrasi pada gambar ditunjukkan bahwa nilai fold adalah 6-fold yang akan dilakukan proses pengujian sejumlah nilai fold tersebut dengan fold ke 1 akan dijadikan data uji dan fold (k-1) dijadikan sebagai data latih. Dengan demikian masing-masing data sampel akan menjadi sebagai data latih dan data uji secara bergantian berdasarkan banyaknya fold yang ditentukan. Hal ini bertujuan untuk meminimalkan nilai akurasi yang dihasilkan oleh faktor kebetulan.

HASIL DAN PEMBAHASAN

Hasil penelitian dan pembahasan penerapan Data Mining dengan menggunakan teknik data mining classification. Tahapan yang digunakan untuk memperoleh hasil penelitian ini yaitu standar proses Data Mining model CRISP-DM (Cross Industry Standard Process for Data Mining) yang dikembangkan tahun 1996 oleh analis dari beberapa industri menetapkan sebagai proses standar strategi pemecahan masalah Data Mining untuk menemukan informasi mengenai nasabah yang akan memperbaharui polisnya.

Atribut yang digunakan seperti Usia, Jumlah anggota keluarga, Sum insured¸ Klaim frekuensi, Pilihan paket, Penyebab meninggal, dan Renewal merupakan data yang berasal dari beberapa tabel yang berbeda pada database pemegang polis dari sistem aplikasi, bahkan beberapa atribut seperti usia, jumlah anggota keluarga, sum insured, klaim frekuensi tidak secara langsung tersimpan dalam basis data, sehingga perlu dibuat suatu fungsi untuk menghitung nilai atribut tersebut. Data yang bersumber dari sistem aplikasi merupakan data pemegang polis yang menggambarkan kondisi nyata yang dimiliki oleh nasabah tersebut. Sumber data atribut pemegang polis yang digunakan berasal dari beberapa tabel yaitu sebagai berikut.

1.Atribut usia diperoleh dari hasil pengurangan tanggal pada kolom tgl_pengajuan dengan kolom tgl_lahir_1 di tabel tm_asri_pengajuan serta tabel tm_asri_pengajuan_expired.
2.Atribut jumlah anggota keluarga diperoleh dari kolom jml_tertanggung pada tabel tm_asri_pengajuan dan tabel tm_asri_pengajuan_expired.
3.Atribut sum insured diperoleh dari kolom total_up pada tabel view v_laporan_aktuaria.
4.Atribut klaim frekuensi diperoleh dari kolom jml_tertanggung_klaim pada tabel view v_laporan_aktuaria.
5.Atribut pilihan paket diperoleh dari kolom pilihan_paket pada tabel tm_asri_premi dan tm_asri_pertanggungan.
6.Atribut penyebab meninggal diperoleh dari kolom penyebab_meninggal pada tabel tm_asri_klaim.
7.Atribut renewal diperoleh dari kolom no_kk pada tabel tm_asri_pengajuan_expired dengan menggabungkan kolom no_kk pada tabel tm_asri_pengajuan.

Train set dan Test set yang akan digunakan adalah data polis klaim. Untuk data polis klaim diambil dari data pemegang polis yang sudah aktifasi polisnya serta data pemegang polis yang sudah kadaluarsa dimana masa aktifnya sudah lewat dari 1 tahun.

Dari seluruh total data pada tabel laporan aktuaria yang berjumlah 445.425 data polis yang terdiri dari 5.368 data polis yang melakukan pengajuan klaim dan 440.057 data polis yang tidak melakukan pengajuan klaim. Sedangkan dalam penelitian ini hanya akan menggunakan 11.604 data polis yang melakukan pengajuan klaim maupun yang tidak melakukan pengajuan klaim.

Karena, setiap polis yang melakukan pembaharuan ataupun yang tidak tentunya memiliki kriteria yang berbeda, maka untuk penelitian ini data yang digunakan untuk memprediksi dalam pembaharuan polis / renewal adalah dengan melihat atribut pendukung dalam melakukan penentuan metode yang tepat dalam penerapan prediksi pembaharuan polis.

Pembangunan model klasifikasi dengan menggunakan C4.5 dimulai dengan memisahkan data latih dan data target. Pembelajaran yang dilakukan untuk menghasilkan model klasifikasi dengan metode C.45 dalam penelitian ini menggunakan bahasa pemprograman java dengan API dari weka.jar. Berikut adalah potongan program untuk menghasilkan model klasifikasi dengan metode C4.5.

KESIMPULAN

Berdasarkan pembahasan yang telah dijelaskan sebelumnya, maka dalam penelitian model prediksi pembaharuan polis asuransi jiwa dengan model Heterogeneous Multiple Classifiers, dapat diambil kesimpulan model Heterogeneous Multiple Classifiers untuk data train menghasilkan nilai akurasi sebesar 94.61%, precision sebesar 95.20%, recall sebesar 94.60% dan f-measure

MEANS (Media Informasi Analisa dan Sistem)
p-ISSN: 2548-6985, e-ISSN:2599-3089

sebesar 94.60%. Sedangkan untuk data test menghasilkan nilai akurasi sebesar 89.00%, precision sebesar 90.30%, recall sebesar 89.00% dan f-measure sebesar 95.60%. Hasil ini lebih tinggi dari hasil yang dicapai oleh model prediksi berbasis Single Classifiers (C4.5 , KNN dan Naïve Bayes) yang digunakan dalam membangun model prediksi berbasis Heterogeneous Multiple Classifiers.

UCAPAN TERIMAKASIH

Penulis mengucapkan terimakasih kepada LRPM dan Prodi Teknik Informatika Universitas Dian Nusantara atas sarana dan dukungan yang diberikan untuk dapat menyelesaikan penelitian ini, semoga penelitian ini dapat bermanfaat khususnya didunia pendidika agar dapat memajukan system khusunya dibidang IT.

DAFTAR PUSTAKA

[1] Avegad, J. 2019. Data Mining Klasifikasi Untuk Memprediksi Status Keberlanjutan Polis
[2] Bashir, S, dkk. 2014. Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble
[3] Chaira, LM, dkk. 2016. Pemilihan Jenis Asuransi Berdasarkan Demografi Calon Pemegang Polis Dengan Metode Naïve Bayes Classifie
[4] Santosa, B. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu
[5] Fadhillah, M., Sari, H. L. ., & Elfianty, L. . (2023). Analisis dan Penerapan Algoritma Naïve Bayes untuk Klasifikasi Penyakit Gingivitis. MEANS (Media Informasi Analisa Dan Sistem), 8(2), 222-226.
[6] Harahap, S. M. ., & Kurniawan, R. . (2024). Analisis Sentimen Komentar Youtube Terhadap Food Vlogger Dengan Menggunakan Metode Naïve Bayes. MEANS (Media Informasi Analisa Dan Sistem), 9(1), 87-96.