 http://dx.doi.org/10.29313/.v0i0.30320

# Klasifikasi Pemegang Polis Menggunakan Metode XGBoost

**Aditya Adam Firdaus*, Aceng Komarudin Mutaqin**

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Islam Bandung, Indonesia.

*itsadityaadam@gmail.com, aceng.k.mutaqin@gmail.com

**Abstract.** Based on data from the Indonesian Police, recorded in 2019 the number of accidents in Indonesia increased by 3 percent from 2018, which is 103,672 accidents. Therefore, the insurance company's step in addressing the uncertainty of the risk that will occur is by choosing an accurate method to predict whether the insured or the policyholder claims the risk that occurs or not. In this study will discuss about one of the methods of gradient decision tree for classification problems. The method is XGBoost. The method will be applied in predicting the classification of claims from Indonesian motor vehicle insurance data in PT insurance companies. X for the 2013 policy. The purpose of this study is to see how accurate the XGBoost method is in classifying claims from Indonesian motor vehicle insurance data at PT insurance companies. X for 2013 policy based on evaluation metrics such as confusion matrix, accuracy and precision. Based on the results, the XGBoost method was able to predict the classification of claims with an average accuracy of 80.87% and precision of 80.48%.

**Keywords: Classification, XGBoost, Evaluation Metrics.**

**Abstrak.** Berdasarkan data dari Kepolisian Republik Indonesia, tercatat pada tahun 2019 jumlah kecelakaan di Indonesia meningkat 3 persen dari 2018, yaitu sebanyak 103.672 kecelakaan. Untuk itu, langkah perusahaan asuransi dalam mengatasi ketidakpastian risiko yang akan terjadi yaitu dengan memilih metode yang akurat untuk memprediksi apakah tertanggung atau pemegang polis mengklaim risiko yang terjadi atau tidak. Dalam penelitian ini akan membahas mengenai salah satu dari metode pohon keputusan gradien untuk permasalahan klasifikasi. Metode tersebut adalah XGBoost. Metode tersebut akan diterapkan dalam memprediksi klasifikasi klaim dari data asuransi kendaraan bermotor Indonesia di perusahaan asuransi PT. X untuk polis tahun 2013. Tujuan dari hasil penelitian ini yaitu melihat seberapa akurat metode XGBoost dalam mengklasifikasikan klaim dari data asuransi kendaraan bermotor Indonesia di perusahaan asuransi PT. X untuk polis tahun 2013 berdasarkan *evaluation metrics* seperti *confusion matrix*, akurasi dan presisi. Berdasarkan hasil yang didapati, metode XGBoost mampu memprediksi klasifikasi klaim dengan rata-rata akurasi sebesar 80,87% dan presisi sebesar 80,48%.

**Kata Kunci: Classification, XGBoost, Evaluation Metrics.**

## 1. Pendahuluan

Asuransi adalah perjanjian antara dua pihak, yaitu perusahaan asuransi dan pemegang polis, yang menjadi dasar bagi penerimaan premi oleh perusahaan asuransi sebagai imbalan untuk memberikan penggantian kepada tertanggung atau pemegang polis, memberikan pembayaran yang didasarkan pada meninggal atau hidupnya tertanggung atau pemegang polis [1].

Dalam asuransi kendaraan bermotor, ada dua jenis perlindungan untuk asuransi kendaraan bermotor, yaitu *Total Loss Only* (TLO) dan *Comprehensive* (komprehensif) [2]. *Comprehensive* disebabkan oleh tabrakan, dsb. Sedangkan, TLO serupa dengan *Comprehensive*, jika kerugian atau kerusakan kendaraan sudah mencapai 75% dari harga pasar kendaraan yang menjadi tanggungan.

Pada tahun 2019 jumlah kecelakaan di Indonesia meningkat 3 persen dari 2018, yaitu sebanyak 103.672 kecelakaan [3]. Untuk itu, langkah perusahaan asuransi dalam mengatasi ketidakpastian risiko yang akan terjadi di masa yang akan datang, perusahaan asuransi harus bersiap-siap dalam memilih metode yang akurat untuk memprediksi apakah tertanggung atau pemegang polis mengklaim risiko yang terjadi atau tidak. Dalam memprediksi klaim tersebut, dapat menggunakan salah satu teknik data mining yaitu pembelajaran mesin dengan metode pengklasifikasian.

Sebuah penelitian yang membahas tentang membangun model yang tepat untuk memprediksi klaim asuransi mobil dengan memfokuskan pada metode statistika lanjutan dan algoritma pembelajaran mesin [4]. Dalam penelitiannya, membandingkan hasil klasifikasi biner yakni klaim atau tidak klaim dari empat metode yakni *Artificial Neural Network* (ANN), *Decision Tree* (DT), *Naïve Bayes classifiers*, dan *eXtreme Gradient Boosting* (XGBoost). Hasil penelitiannya menunjukkan bahwa metode XGBoost mencapai akurasi dan nilai *Receiver Operator Characteristic* (ROC) yang terbaik dari keempat metode, yaitu sebesar 92,53% dan 0,986.

Berdasarkan hasil penelitian diatas, dalam penelitian ini akan dilakukan prediksi klasifikasi pemegang polis menggunakan metode XGBoost terhadap data asuransi kendaraan bermotor di Indonesia.

## 2. Metodologi

Metode yang digunakan dalam penelitian ini yaitu XGBoost. XGBoost (*eXtreme Gradient Boosting*) merupakan implementasi dari pohon keputusan gradien yang dirancang untuk kecepatan dan kinerja [5]. Dalam penggunaannya XGBoost digunakan untuk permasalahan *supervised learning*, yang menggunakan *data training* dimana $x_i$ untuk memprediksi variabel target $y_i$. Secara matematikanya, dalam menulis nilai prediksi pada langkah ke $(t)$ pada $\hat{y}_i{}^{(t)}$ sebagai berikut [6]:

$$\hat{y}_i{}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i{}^{(t-1)} + f_t(x_i) \tag{1}$$

Dalam XGBoost terdapat fungsi obyektif *training loss* dan *regularization*. Fungsi obyektifnya yaitu,

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{2}$$

Dimana, $L$ adalah *training loss function* dan $\Omega$ adalah *regularization*.

*Training loss* yaitu mengukur seberapa prediktif model dengan *data training*. pilihan umum yang sering digunakan yaitu *mean squared error* sebagai berikut:

$$L(\theta) = \sum_{i} (y_i - \hat{y}_i)^2 \tag{3}$$

Kemudian dalam mendefinisikan kompleksitas pada *regularization* sebagai berikut:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

Setelah melihat matematika dibalik algoritma XGBoost dibuat. Dalam mendapatkan nilai *evaluation metrics* seperti *confusion* matrix, akurasi dan presisi dari hasil klasifikasi dengan menggunakan XGBoost terdapat parameter-parameter yang dapat disesuaikan nilainya agar mendapatkan hasil yang baik:

**Tabel 1.** Parameter yang Terdapat pada *eXtreme Gradient Boosting*

| PARAMETER | KETERANGAN |
|---|---|
| ***n_estimators*** | Jumlah pohon (*tree*) |
| ***learning_rate*** | Penyusutan ukuran yang digunakan untuk mencegah *overfitting* |
| ***n_jobs*** | Jumlah *threads* paralel yang digunakan |
| ***gamma*** | Meminimalisir kerugian yang diperlukan untuk membuat partisi lebih lanjut pada *leaf node* pohon |
| ***reg_alpha*** | Regularisasi L1 pada pembobot (alfa / $\alpha$) |
| ***reg_lambda*** | Regularisasi L2 pada pembobot (lambda / $\lambda$) |
| ***random_state*** | *Seed* yang digunakan oleh generator angka acak |

Setelah melihat parameter-parameter yang digunakan dalam XGBoost. Hasil akhir dalam melihat kualitas metodenya baik atau tidak yaitu dengan mengevaluasinya. Untuk evaluasinya menggunakan *evaluation metrics* yakni,

Akurasi, pembagian dari jumlah prediksi benar terhadap total prediksi. Persamaan (5) menyajikan perhitungan dari akurasi [7].

$$Classification\ Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{5}$$

*Confusion matrix*, pencarian ukuran performa dari hasil klasifikasi. Tabel 2 memberikan contoh *confusion matrix* dengan dua klasifikasi yakni benar (*true*) dan salah (*false*) [7].

**Tabel 2.** Confusion Matrix

| | | **Nilai Aktual** | |
|---|---|---|---|
| | | **Positif** | **Negatif** |
| **Nilai Prediksi** | **Positif** | *True Positive* (TP) | *False Positive* (FP) |
| | **Negatif** | *False Negative* (FN) | *True Negative* (TN) |

Presisi, pembagian dari jumlah total contoh positif yang diklasifikasikan bernilai benar dengan jumlah total contoh positif yang diprediksi [7]. Perhitungannya sebagai berikut:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

## 3. Pembahasan dan Diskusi

### Data

Data yang digunakan yakni data sekunder yang diperoleh dari perusahaan asuransi PT. X untuk polis tahun 2013. Data yang diperoleh berisikan informasi polis asuransi kendaraan bermotor kategori 3 (uang pertanggungan >Rp200.000.000,00 s.d. Rp400.000.000,00) dan klaim dari setiap polis pada wilayah 1 (Sumatera dan Kepulauan di sekitarnya).

Variabel yang terdapat dalam data dan digunakan dalam penelitian ini yaitu disajikan pada Tabel 3.

**Tabel 3.** Variabel-variabel Data Asuransi Kendaraan Bermotor Indonesia

| VARIABEL | JENIS DATA | KETERANGAN |
|---|---|---|
| kode_perusahaan | Kategorik | Kode dari sebuah perusahaan |
| kode_polis | Kategorik | Kode internal Perusahaan |
| nomor_rangka | Kategorik | Nomor rangka kendaraan polis |

| | | |
|---|---|---|
| nomor_mesin | Kategorik | Nomor mesin kendaran polis |
| nomor_polisi | Kategorik | Identitas kendaraan polis |
| kode_pertanggungan | Kategorik | Jenis pertanggungan polis |
| kode_penggunaan | Kategorik | Jenis kendaraan yang digunakan |
| kode_wilayah | Kategorik | Alamat wilayah polis |
| tahun_kendaraan | Numerik | Tahun kendaraan polis |
| harga_pertanggungan | Numerik | Taksiran harga pertanggungan |
| merek_kendaraan | Kategorik | Merek kendaraan polis |
| tipe_kendaraan | Kategorik | Tipe dari merek kendaraan polis |
| target | Kategorik | Polis melakukan klaim atau tidak klaim |

Variabel yang akan digunakan dalam pemodelan yakni variabel target akan menjadi variabel respon, untuk variabel kode_perusahaan, kode_pertanggungan, kode_penggunaan, kode_wilayah, tahun_kendaraan, harga_pertanggungan, merek_kendaraan dan tipe_kendaraan sebagai variabel prediktor. Sedangkan sisanya akan digunakan sebagai eksplorasi. Pada variabel kode polis, nomor rangka, nomor mesin, nomor polisi dan kode penggunaan termasuk jenis data kategorik dan masih dalam keadaan *string*, maka perlu dilakukan pengkodean terlebih dahulu, agar bisa digunakan dalam eksplorasi dan pemodelan. Sebagai contoh Tabel 4 menyajikan pengkodean pada variabel kode polis dengan 23.548 kategorik.

**Tabel 4.** Pengkodean Variabel Kode Polis

| SEBELUM PENGKODEAN | SESUDAH PENGKODEAN |
|---|---|
| 000000000359202 | 0 |
| ⋮ | ⋮ |
| ZIPMZ214566811 | 23547 |

**Hasil dan Pembahasan**

Data dalam penelitian ini terdiri dari 33.224 pengamatan dan 13 variabel, kemudian dilakukan *data preprocessing*. Hasilnya pada variabel nomor polisi dideteksi terdapat 1 baris pengamatan yang mengandung *missing values* dan 2 data duplikasi untuk polis yang tidak mengajukan klaim yang terdapat dalam data. Tiga pengamatan tersebut akan dikeluarkan dari data. Hal ini mengakibatkan data yang akan digunakan menjadi 33.221 pengamatan. Selanjutnya data tersebut dilakukan pengkodean untuk bisa dilanjutkan ke dalam eksplorasi dan pemodelan.

Setelah dilakukannya *data preprocessing*, selanjutnya eksplorasi data menampilkan hasil dari tabel frekuensi untuk polis yang melakukan klaim, dimana hasilnya disajikan dalam Tabel 5. Sedangkan tabel frekuensi untuk setiap polis yang tidak melakukan klaim disajikan dalam Tabel 6.

**Tabel 5.** Tabel Frekuensi untuk Polis yang Klaim

| No. | Kode Polis Yang Klaim | Jumlah |
|---|---|---|
| 1 | 2000000056-3445-23584-786-12561 | 15 |
| 2 | 2000000056-9088-23618-824-18939 | 14 |
| ⋮ | ⋮ | ⋮ |
| 7989 | 2000000056-4068-9750-8609-5841 | 1 |
| 7990 | 2000003803-23425-15052-22490-13454 | 1 |
| **Total** | | **14726** |

Penjelasan isi dari Tabel 5 adalah sebagai berikut. Sebagai contoh dari 7.990 polis yang melakukan klaim asuransi kendaraan bermotor di tahun 2013, polis dengan kode klaim 2000000056-3445-23584-786-12561 (kode perusahaan 2000000056, kode polis 3445, nomor rangka 23584, nomor mesin 786, nomor polisi 12561) melakukan klaim sebanyak 15 kali selama

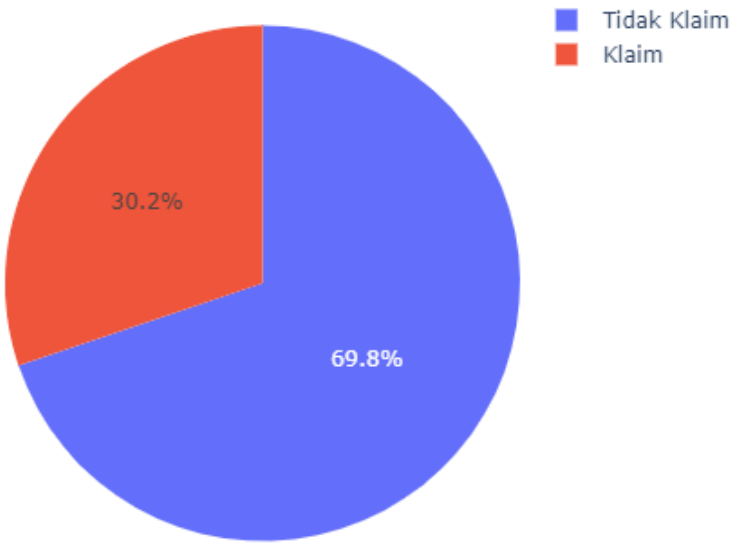
masa asuransinya. Contoh lainnya dapat dijelaskan dengan cara yang sama sebagaimana di atas.

**Tabel 6.** Tabel Frekuensi untuk Polis yang Tidak Klaim

| No. | Kode Polis Yang Tidak Klaim | Jumlah |
|---|---|---|
| 1 | 2000000048-6355-3693-6991-23013 | 1 |
| 2 | 2000003794-6600-10116-5282-10884 | 1 |
| ⋮ | ⋮ | ⋮ |
| 18494 | 2000000103-1525-5233-8156-15071 | 1 |
| 18495 | 2000005345-19450-147-4231-14569 | 1 |
| **Total** | | **18495** |

Penjelasan isi dari Tabel 6 adalah sebagai berikut. Dari data polis asuransi kendaraan bermotor tahun 2013, terdapat 18.495 polis yang tidak mengajukan klaim di masa asuransinya.

Gambar 1 menyajikan diagram lingkaran persentase polis yang melakukan klaim dan tidak melakukan klaim. Berdasarkan Gambar 1 terlihat bahwa ada sekitar 69,8% polis tidak mengajukan klaim di masa asuransinya, sisanya yaitu sekitar 30,2% polis mengajukan klaim di masa asuransinya.




**Gambar 1.** Persentase Polis yang Klaim dan Tidak Klaim

Setelah eksplorasi data, dilanjutkan dengan *splitting*. Dimana data polis yang totalnya ada 26.485 polis akan dilakukan *cross validation* (memvalidasi berulang kali dimana dataset dibagi menjadi banyak *subset* (iterasi) *training* dan *testing*). *Cross validation* yang digunakan yaitu *stratified k-fold* dengan $k = 5$. Karena nilai $k = 5$, proses ini akan dilakukan pengulangan selama 5 kali dengan proporsi 80% *data training* atau 21.188 polis dan 20% *data testing* atau 5.297 polis.

Untuk menjalankan proses pemodelannya, terlebih dahulu menentukkan nilai parameter yang disesuaikan ke dalam model. Tabel 6 menyajikan hasil nilai parameter yang telah disesuaikan.

**Tabel 6.** Hasil Nilai Parameter Digunakan Dalam Model

| Parameter | Nilai Parameter |
|---|---|
| *n_estimators* | 1.000 |
| *learning_rate* | 0,001 |
| *n_jobs* | -1 |
| *gamma* | 0,0 |
| *reg_alpha* | 2 |
| *reg_lambda* | 10 |
| *random_state* | 0 |

Penjelasan isi dari Tabel 6 adalah sebagai berikut. Jumlah pohon (*n_estimators*) yang digunakan dalam mencari model terbaik yakni sebanyak 1.000. Dalam penyusutan ukuran untuk mencegah overfitting (*learning_rate*) yakni sebesar 0,001. Jumlah *threads* paralel yang digunakan (*n_jobs*) yakni -1 artinya menggunakan semua inti CPU pada sistem dalam proses pembelajaran.

Setelah menentukan nilai parameter, maka dapat dilakukan pemodelan dari kelima iterasi dan diperoleh hasil *output*nya. Tabel 7 menyajikan hasil *confusion matrix* dari 5 iterasi. Sedangkan Tabel 8 menyajikan hasil akurasi dan presisi dari 5 iterasi.

**Tabel 7.** *Confusion Matrix* dari Pemodelan XGBoost

| | | | | Aktual | |
|---|---|---|---|---|---|
| | | | | Tidak Klaim | Klaim |
| Iterasi | 1 | Prediksi | Tidak Klaim | 3.533 | 845 |
| | | | Klaim | 166 | 753 |
| | 2 | | Tidak Klaim | 3.565 | 881 |
| | | | Klaim | 134 | 717 |
| | 3 | | Tidak Klaim | 3.526 | 854 |
| | | | Klaim | 173 | 744 |
| | 4 | | Tidak Klaim | 3.551 | 903 |
| | | | Klaim | 148 | 695 |
| | 5 | | Tidak Klaim | 3.556 | 819 |
| | | | Klaim | 143 | 779 |

Penjelasan isi dari Tabel 7 adalah sebagai berikut. Sebagai contoh dari 3.699 polis yang tidak melakukan klaim, model memprediksi polis tersebut dengan tepat sebanyak 3.533, sisanya 166 polis diprediksi tidak tepat. Sedangkan dari 1.598 polis yang melakukan klaim, model memprediksi polis tersebut dengan tepat sebanyak 753, sisanya 845 polis diprediksi tidak tepat. Contoh lainnya dapat dijelaskan dengan cara yang sama sebagaimana di atas.

**Tabel 8.** Akurasi dan Presisi dari Pemodelan XGBoost

| Iterasi | 1 | 2 | 3 | 4 | 5 | Rata-rata |
|---|---|---|---|---|---|---|
| **Akurasi** | 80,91% | 80,84% | 80,61% | 80,16% | 81,84% | 80,87% |
| **Presisi** | 80,70% | 80,18% | 80,5% | 79,73% | 81,28% | 80,48% |

Berdasarkan Tabel 8 terlihat bahwa metode XGBoost dapat mengklasifikasikan klaim dengan akurat sebesar 80,87%. Sedangkan dalam hal presisi klasifikasi klaimnya sebesar 80,48%.

## 4. Kesimpulan

Dalam penelitian ini dibahas klasifikasi pemegang polis menggunakan metode XGBoost berdasarkan data asuransi kendaraan bermotor kategori 3 wilayah 1 untuk polis tahun 2013 di Indonesia. Berdasarkan hasil penelitian didapatkan bahwa metode XGBoost dapat mengklasifikasikan klaim dengan akurat sebesar 80,87%. Sedangkan dalam hal presisi klasifikasi klaimnya sebesar 80,84%. Hal ini membuktikan bahwa metode XGBoost dapat diterapkan dengan baik dalam hal klasifikasi pemegang polis terhadap data asuransi kendaraan bermotor kategori 3 wilayah 1 untuk polis tahun 2013 di Indonesia.

**Acknowledge**

We gratefully thank to Seminar Penelitian Sivitas Akademika UNISBA (SpeSIA UNISBA) for organizing this article until the end.

## Daftar Pustaka

[1] Republik Indonesia. 2014. Undang-Undang Republik Indonesia Nomor 40 Tahun 2014 tentang Perasuransian. Diambil kembali dari https://www.ojk.go.id/Files/201506/1UU402014Perasuransian_1433758676.pdf

[2] Kartini, N., Sunendiari, S., & Mutaqin, A. K. (2018). Penentuan Distribusi Kerugian Agregat Tertanggung Asuransi Kendaraan Bermotor di Indonesia Menggunakan Metode Fast Fourier Transform. Prosiding Statistika, Vol. 4, No. 1, 18-25.

[3] PT. Kompas Cyber Media (Kompas Gramedia Digital Group). 2019. Polri Sebut Jumlah Kecelakaan Lalu Lintas Meningkat pada 2019. Diambil kembali dari Kompas.com: https://nasional.kompas.com/read/2019/12/28/10355741/polri-sebut-jumlah-kecelakaan-lalu-lintas-meningkat-pada-2019

[4] Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. 2020. A Proposed Model to Predict Auto Insurance Claims Using Machine Learning Techniques. Vol. 98, 22.

[5] Brownlee, J. 2018. XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn (1.10 ed.). Machine Learning Mastery.

[6] XGBoost Developers. (2021, Juni 1). XGBoost Release 1.5.0-dev. Diambil kembali dari XGBoost: https://buildmedia.readthedocs.org/media/pdf/xgboost/latest/xgboost.pdf

[7] Yunus, M. (2020, Januari 12). #3 Machine Learning Evaluation. Diambil kembali dari Medium: https://yunusmuhammad007.medium.com/3-machine-learning-evaluation-239426e3319e

[8] Shofwani Sheila Ghazia, Kudus Abdul. (2021). *Penentuan Kriteria Pengunjung dalam Pemilihan Green Hotel di Kota Bandung Menggunakan Metode Discrete Choice Experiment dengan Desain Choice Sets Kombinatorial*. Jurnal Riset Statistika, 1(1), 1-9.