Volume 10 Number 1. Mei 2020 Page. 57-64
Journal Homepage : http://teknois.stikombinaniaga.ac.id/index.php/JBS
DOI Link : http://doi.org/10.36350/jbs.v10i1

Article DOI : 10.36350/jbs.v10i1.78
Received: Maret; Accepted : April; Published : Mei

Resampling Neural Network untuk Penanganan Class Imbalance pada Prediksi Klaim Asuransi

Hudori
Teknik Informatika/STIKOM Binaniga
Email: i_am_dhori@yahoo.com

ABSTRACT

Neural Network algorithm has advantages of the calculation accuracy that is better than another algorithm because calculation process is done over and over again so that it takes a longer time in processing or training the data. However, this algorithm is also very sensitive against the dataset that has a very unbalanced class. Handling the class imbalance that occurs in the dataset can be overcome by resampling technique. This paper compares the three methods of resampling to handle an imbalance of class applied to the algorithm of Neural Network as one of the mining data algorithms for the prediction of Bodily Injury Claim of Passengers Based on the Characteristics of the Insured Vehicle. This method is built and tested using the real data transactions from a leading insurance company in the world asking a crowdsourcing company to organize a competition of prediction model construction of Passenger Bodily Injury Claim Based on the Characteristics of the Insured Vehicle. In General, transaction data, this data is also still having high dimensional, heterogeneous properties and empty value on some variable.

Keywords*: Neural Network; Imbalance Class; Resampling; Claim Prediction.*

ABSTRAK

Algoritma Neural Network memiliki kelebihan akan tingkat akurasi perhitungan yang lebih baik dari algoritma lain karena proses perhitungan dilakukan berulang-ulang sehingga membutuhkan waktu yang lebih lama dalam memproses atau men-training data. Namun algoritma ini juga sangat sensitive terhadap dataset yang memiliki class yang sangat tidak seimbang. Penanganan ketidakseimbangan kelas (class imbalance) yang terjadi pada dataset dapat diatasi dengan teknik resampling. Tulisan ini membandingkan tiga metode resampling untuk menangani ketidak-seimbangan class yang diterapkan pada algoritma Neural Network sebagai salah satu algoritma data mining untuk prediksi Claim Kecelakaan Diri Penumpang Berdasarkan Karakteristik Kendaraan Tertanggung. Metode ini dibangun dan diuji menggunakan data transaksi yang real dari sebuah perusahaan asuransi terkemuka di dunia yang meminta sebuah perusahaan crowdsourching untuk menyelenggarakan kompetisi pembangunan model prediksi Claim Kecelakaan Diri Penumpang Berdasarkan Karakteristik Kendaraan Tertanggung. Pada umumnya data transaksi, data ini juga masih memiliki sifat berdimensi tinggi, heterogen dan nilai kosong pada beberapa variable.

Kata Kunci*: Neural Network; Ketiakseimbangan Kelas; Resampling; Prediksi Klaim.*

A. PENDAHULUAN

Volume 10 Number 1. Mei 2020 Page. 57-64
Journal Homepage : http://teknois.stikombinaniaga.ac.id/index.php/JBS
DOI Link : http://doi.org/10.36350/jbs.v10i1

1. Latar Belakang

Data *claim* pada perusahaan asuransi biasanya memiliki sifat yang *high class imbalance*, karena perusahaan asuransi memang mengharapkan rasio klaim yang rendah pada *frequency* maupun severity-nya terhadap polis asuransi yang diterbitkannya. Dengan demikian, semakin rendah rasio frequency dan severity claim terhadap jumlah produksi polis maka semakin tinggi class imbalance pada data tersebut. Hal ini akan menyebabkan fenomena accuracy paradox, dimana accuracy yang tinggi tidak mencerminkan baiknya metode prediksi.

Penanganan ketidakseimbangan kelas (class imbalance) yang terjadi pada dataset dapat diatasi dengan teknik resampling (Cateni, Colla, & Vannucci, 2014). Teknik resampling akan memanipulasi kelas distribusi dengan memperbaiki data latih sehingga didapatkan kelas yang seimbang. Terdapat beberapa teknik resampling antara lain: random over sampling, random under sampling, directed over sampling, directed under sampling, kombinasi dari over sampling dan under sampling, dan Synthetic Minority Oversampling Technique (SMOTE) (Cateni et al., 2014; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Metode cluster-based under-sampling yang diterapkan oleh Show Jane Yen juga merupakan metode resampling yang menerapakan teknik under sampling menggunakan neural network untuk mengklasifikasikan dataset (Yen & Lee, 2009). Metode Similarity-based UnderSampling and Normal Distribution-based Oversampling (SUNDO) adalah metode resampling yang mengkombinasikan teknik under sampling dan over sampling sehingga menghasilkan kinerja dataset yang lebih baik (Cateni et al., 2014). Dengan menggunakan metode resampling pada pengolahan data, maka akan didapatkan keseimbangan kelas yang merata.

Penelitian sebelumnya (Pozollo 2011) menyimpulan bahwa penerapan tehnik Decision Tree menghasilkan nilai akurasi dan computing cost yang lebih baik ketika dibandingkan dengan tehnik Support Vector Machine (SVM), Naive Bayes, Random Forest, Linear Discriminant Analysis, Neural Network dan K-Nearest Neighborhood. Namun akurasi tidak selalu mencerminkan baiknya model prediksi, sehingga perlu kajian lebih lanjut terhadap kasus ini. Karena data yang digunakan memiliki kelas target yang sangat tidak seimbang (99.45% : 0.55%).

Penelitian ini mencoba mengukur kinerja tehnik Neural Network yang didahului dengan melakukan langkah preprocessing pada data melalui tehnik Resampling. Diharapkan dengan tehnik tersebut mampu menaikan kinerja algoritma Neural Network untuk menangani ketidaksimbangan kelas pada data berskala besar.

2. Identifikasi Masalah

Dari uraian latar belakang diatas, maka identifikasi masalah dalam penelitian ini adalah Bagaiamana metode *Resampling Neural Network* dalam menguji ketidakseimbangan kelas (*class imbalance*) dataset skala besar untuk meningkatkan akurasi algoritma *Neural Network* pada prediksi Klaim Asuransi Kecelakaan Diri Penumpang Berdasarkan Karakteristik Kendaraan yang Diasuransikan?

3. Tujuan

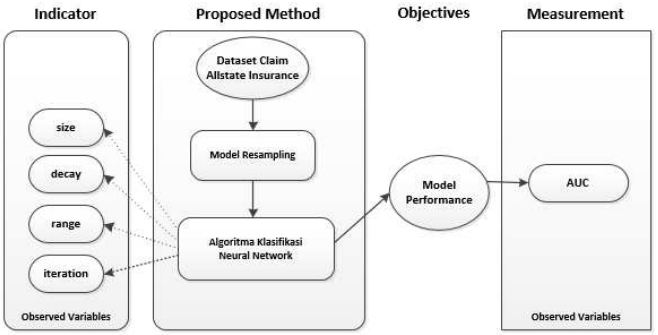
Adapun tujuan dari penelitian ini adalah untuk mendapatkan klasifikasi klaim atau tidak klaim pada prediksi polis Asuransi Kecelakaan Diri Penumpang.

B. METODE

Untuk mengevaluasi hasil klasifikasi, digunakan pengukuran (measurement) dengan Area Under the ROC (*Receiver Operating Characteristic*) *Curve* (AUC). Kerangka pemikiran penelitian ini dapat dilihat pada Gambar 1.

Menurut (Dawson, 2009) terdapat empat metode penelitian yang umum digunakan, yaitu: penelitian langsung, eksperimen, study kasus dan survey. Penelitian ini menggunakan metode eksperimen, yaitu penelitian yang melibatkan penyelidikan pada perameter, variabel atau perbaikan metode tergantung dari penelitinya dan menggunakan data tes yang dikendalikan oleh peneliti itu sendiri.

Volume 10 Number 1. Mei 2020 Page. 57-64
Journal Homepage : http://teknois.stikombinaniaga.ac.id/index.php/JBS
DOI Link : http://doi.org/10.36350/jbs.v10i1



Gambar 1. Kerangka Pemikiran

Gambar 2. Metode Penelitian

1. Pengumpulan Data

Data yang digunakan pada penelitian ini diambil dari kompetisi yang diadakan oleh **Kaggle** dan disponsori oleh **Allstate Insurance** pada tahun 2011 mengenai Prediksi Pembayaran Klaim Asuransi Kecelakaan Diri Penumpang Berdasarkan Karakteristik Kendaraan yang Diasuransikan (*Predicting Bodily Injury Liability Insurance claim payments based on the characteristics of the insured's vehicle*). Data ini bersifat bebas dan dapat diakses melalui situs kaggle (www.kaggle.com).

2. Pengolahan Data Awal

Untuk mempercepat waktu komputasi dan terbatasnya peralatan yang digunakan penulis, maka data yang digunakan dalam penelitian ini dimabil dan terbagi dalam 11 (sebelas) bagian, yang masing-masing bagian diambil dari data claim Allstate Insurance yang sudah melalui proses data cleansing, data transform dan data normalization.

Tabel 1. Data Source Summary

Data Source	Total Variable	Total Data	Claim		No Claim	
			Data	Ratio (%)	Data	Ratio (%)
Train Set	32	13,184,290	95,605	0.73	13,088,685.00	99.27
Test Set	31	4,314,865	0	-	4,314,865.00	100.00
Merging Set	32	17,499,155	95,605	0.55	17,403,550.00	99.45
Sample Set	***24***	***2,155,948***	***74,479***	***3.45***	***2,081,469.00***	***96.55***

Pada penelitian ini, penulis menggunakan metode Slovin (Riduwan 2005:65) dalam pengambilan 11 (sebelas) bagian sample data. Dengan alasan lebih mudah dan sederhana.

$$\mathbf{n = N/N(d)^2 + 1}$$

dimana :

n = sampel

N = populasi

d = nilai presisi 95% atau sig. = 0,05.

Dengan metode di atas, Secara bergantian data diambil secara acak sejumlah 9.954 data, untuk kemudian diulang prosesnya dengan mekanisme yang sama sebanyak sebelas kali. Dimana tidak ada data yang diambil lebih dari satu kali di setiap bagian dari sebelas bagian yang akan dijadikan sample untuk eksperimen.

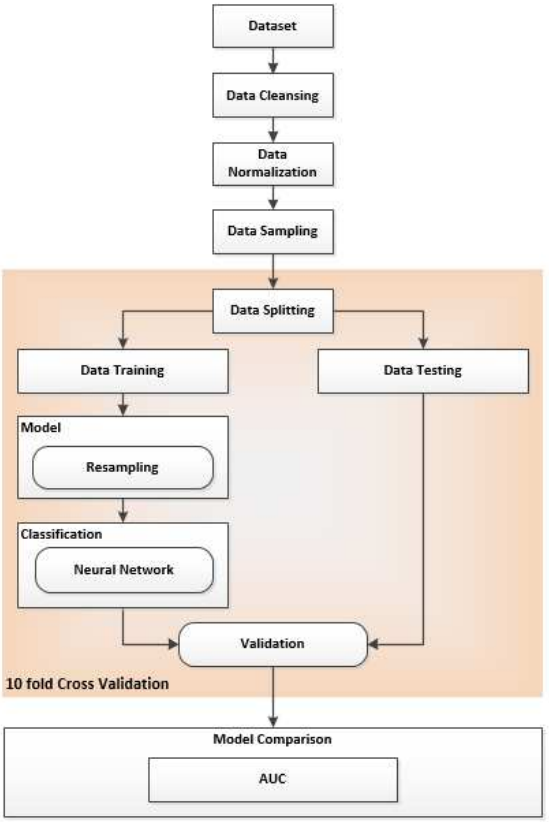
Volume 10 Number 1. Mei 2020 Page. 57-64
Journal Homepage : http://teknois.stikombinaniaga.ac.id/index.php/JBS
DOI Link : http://doi.org/10.36350/jbs.v10i1

Tabel 2. Data Summary pada data sample

Dataset	Row	Incurred			
		Y	Ratio	N	Ratio
dc01	9,954	331	3.33%	9,623	96.67%
dc02	9,954	348	3.50%	9,606	96.50%
dc03	9,954	344	3.46%	9,610	96.54%
dc04	9,954	333	3.35%	9,621	96.65%
dc05	9,954	360	3.62%	9,594	96.38%
dc06	9,954	340	3.42%	9,614	96.58%
dc07	9,954	366	3.68%	9,588	96.32%
dc08	9,954	357	3.59%	9,597	96.41%
dc09	9,954	347	3.49%	9,607	96.51%
dc10	9,954	318	3.19%	9,636	96.81%
dc11	9,954	328	3.30%	9,626	96.70%

3. Metode yang Diusulkan

Metode yang diusulkan yaitu penerapan resampling Neural Network untuk prediksi Klaim Asuransi Kecelakaan Diri Penumpang Berdasarkan Karakteristik Kendaraan yang Diasuransikan. Dimulai dari pembagian dataset dengan metode 10 cross validation yaitu data training dan data testing, kemudian data trainig diproses dengan metode resampling Neural Network, sehingga menghasilkan model evaluasi yang diukur dengan Friedmen test dan Area Under Curve (AUC), dapat dilihat pada Gambar 3.



Gambar 3. Metode yang diusulkan

4. Eksperimen dan pengujian model (Method Test and Experiment)

Proses eksperimen dan pengujian metode dalam penelitian ini menggunakan algoritma open source dan fungsi-fungsi terkait yang ada di aplikasi R, yaitu sebuah perangkat lunak terpopuler untuk data mining dan pengolahan data statistik yang didistribusikan secara gratis.
Adapun tahapan-tahapan eksperimen dan pengujian model pada penelitian ini adalah sebagai berikut:

a. Menyiapkan dataset untuk eksperiment yang sudah di download.
b. Memilih parameter filter dataset yang akan diuji.
c. Melakukan training dan testing pada dataset terhadap model Neural Network dan mencatat hasil perhitungan akurasi, AUC, F-Meansure, G-Mean, PPV dan NPV.

d. Melakukan training dan testing pada dataset terhadap model resampling Neural Network dan mencatat hasil perhitungan akurasi, AUC, F-Meansure, G-Mean, PPV dan NPV.
e. Melakukan komparasi hasil AUC pada model dan menguji beda dengan t-Test.

Dalam penelitian yang dilakukan ini menggunakan komputer untuk melakukan proses perhitungan terhadap model yang diusulkan.

5. Evaluasi dan Validasi Hasil (Result Evalutation and Validation)

Tabel confusion matrix digunakan untuk melakukan pengukuran kinerja model. Kinerja yang diukur termasuk akurasi secara umum, akurasi dalam memprediksi kelas monoritas, dan *Area Under the ROC Curve* (AUC). *Confusion matrix* diperoleh dari proses validasi menggunakan *10-fold cross validation*, sehingga model yang terbentuk dapat langsung diuji dengan melakukan 10 kali pengujian.

Kinerja model yang diperoleh digunakan untuk membandingkan antara sembilan model yang menjadi objek penelitian ini. Untuk melihat kualitas model yang didapatkan, nilai AUC dapat dijadikan ukuran untuk melihat model yang terbentuk. Kurva ROC dapat digunakan untuk mendapatkan nilai AUC, dimana nilai AUC digunakan untuk menentukan klasifikasi pengujian diagnostik.

Pedoman umum untuk mengklasifikasikan keakuratan pengujian diagnostik menggunakan AUC dapat dilihat pada sistem tradisional yang dijabarkan oleh Gorunescu (Wahono, Herman, & Ahmad, 2011), disajikan pada Tabel 3.

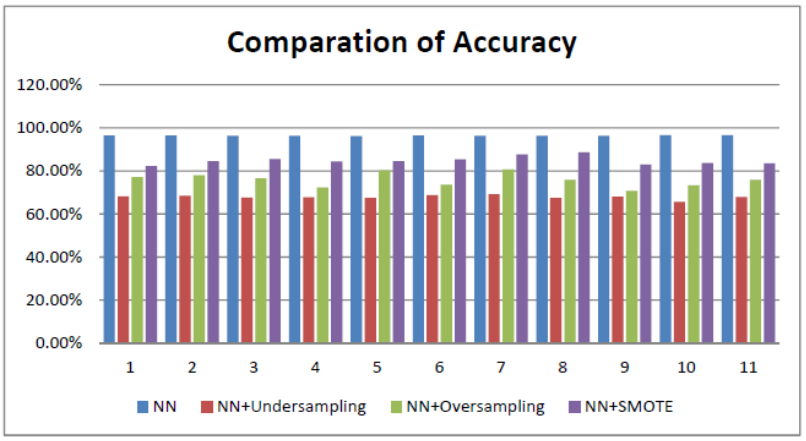
Table 3.3 Klasifikasi AUC

Performance	Klasifikasi
0,90 – 1,00	Paling baik
0,80 – 0,90	Baik
0,70 – 0,80	Adil atau sama
0,60 – 0,70	Rendah
0,50 – 0,60	Gagal

C. HASIL DAN PEMBAHASAN

1. Akurasi

Pada gambar 4. di bawah adalah perbandingan nilai Akurasi dari NN, NN+Undersampling, NN+Oversampling, NN+SMOTE dari 11 (sebelas) dataset yang menjadi dataset training dan dataset test.



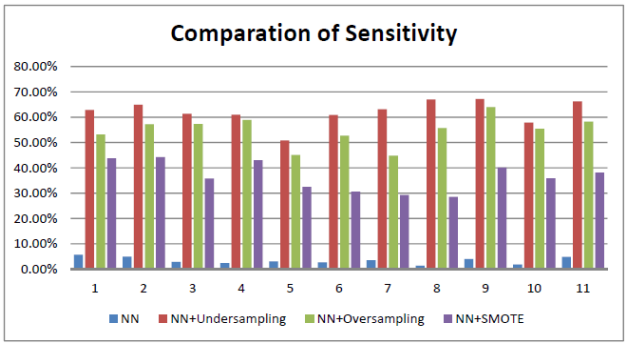
Gambar 4. Perbandingan Akurasi NN, NN+Undersampling, NN+Oversampling, NN+SMOTE

Pada hasil visualisasi di atas dapat ditarik kesimpulan dari percobaan terhadap sebelas dataset dengan menggunakan algoritma NN, NN+Undersampling, NN+Oversampling, NN+SMOTE bahwa kinerja algoritma NN lebih unggul dibandingkan dengan algoritma NN menggunakan Resamping (NN+Undersampling, NN+Oversampling, NN+SMOTE). Akurasi paling rendah adalah metode NN+ Undersampling.

2. Sensitivitas

Pada gambar 5 di bawah adalah perbandingan nilai Sensitivitas dari NN, NN+Undersampling, NN+Oversampling, NN+SMOTE SMOTE dari 11 (sebelas) dataset yang menjadi dataset training dan dataset test.

Volume 10 Number 1. Mei 2020 Page. 57-64
Journal Homepage : http://teknois.stikombinaniaga.ac.id/index.php/JBS
DOI Link : http://doi.org/10.36350/jbs.v10i1

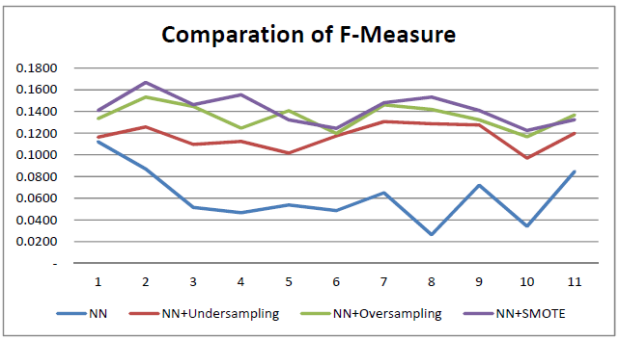


Gambar 5 Perbandingan Sensitivitas NN, NN+Undersampling, NN+Oversampling, NN+SMOTE

Pada hasil visualisasi di atas dapat ditarik kesimpulan dari percobaan terhadap sebelas dataset dengan menggunakan algoritma NN, NN+Undersampling, NN+Oversampling, NN+SMOTE bahwa kinerja algoritma NN+Undersampling lebih unggul dibandingkan dengan algoritma NN dan Algoritma NN+Resampling yang lain (NN+Oversampling, NN+SMOTE). Sensitivitas paling rendah adalah metode NN tanpa Resampling.

3. F-Measure

Pada gambar 6 di bawah adalah perbandingan nilai F-Measure dari NN, NN+Undersampling, NN+Oversampling, NN+SMOTE SMOTE dari 11 (sebelas) dataset yang menjadi dataset training dan dataset test.

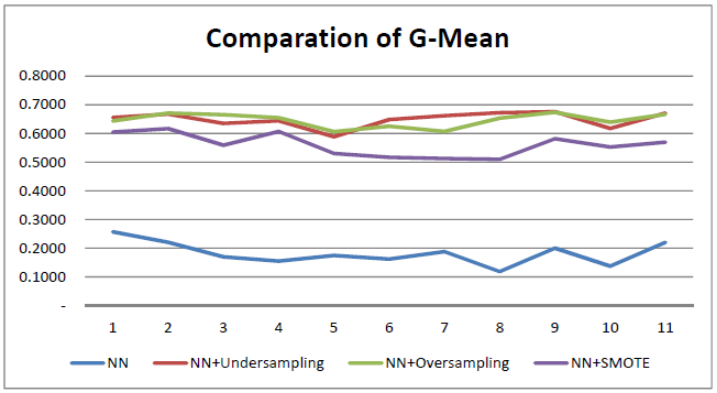


Gambar 6 Perbandingan F-Measure NN, NN+Undersampling, NN+Oversampling, NN+SMOTE

Pada hasil visualisasi di atas dapat ditarik kesimpulan dari percobaan terhadap sebelas dataset dengan menggunakan algoritma NN, NN+Undersampling, NN+Oversampling, NN+SMOTE bahwa kinerja algoritma NN+SMOTE lebih unggul dibandingkan dengan algoritma NN dan Algoritma NN+Resampling yang lain (NN+Undersampling , NN+Oversampling). F-Measure paling rendah adalah metode NN tanpa Resampling.

4. G-Mean

Pada gambar 7 di bawah adalah perbandingan nilai G-Mean dari NN, NN+Undersampling, NN+Oversampling, NN+SMOTE SMOTE dari 11 (sebelas) dataset yang menjadi dataset training dan dataset test.



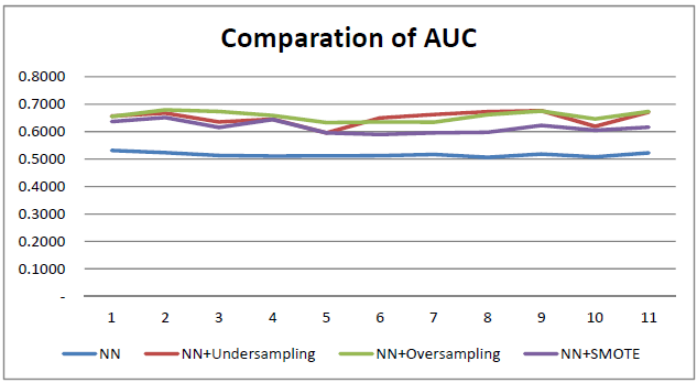
Gambar 7. Perbandingan G-Mean NN, NN+Undersampling, NN+Oversampling, NN+SMOTE

Volume 10 Number 1. Mei 2020 Page. 57-64
Journal Homepage : http://teknois.stikombinaniaga.ac.id/index.php/JBS
DOI Link : http://doi.org/10.36350/jbs.v10i1

Pada hasil visualisasi di atas dapat ditarik kesimpulan dari percobaan terhadap sebelas dataset dengan menggunakan algoritma NN, NN+Undersampling, NN+Oversampling, NN+SMOTE bahwa kinerja algoritma NN+Undersampling dan NN+Oversampling lebih unggul dibandingkan dengan algoritma NN dan Algoritma NN+SMOTE. G-Mean paling rendah adalah metode NN tanpa Resampling.

5. AUC

Pada gambar 8 di bawah adalah perbandingan nilai AUC dari NN, NN+Undersampling, NN+Oversampling, NN+SMOTE SMOTE dari 11 (sebelas) dataset yang menjadi dataset training dan dataset test.



Gambar 8. Perbandingan AUC NN, NN+Undersampling, NN+Oversampling, NN+SMOTE

Pada hasil visualisasi di atas dapat ditarik kesimpulan dari percobaan terhadap sebelas dataset dengan menggunakan algoritma NN, NN+Undersampling, NN+Oversampling, NN+SMOTE bahwa kinerja algoritma NN+Undersampling dan NN+Oversampling lebih unggul dibandingkan dengan algoritma NN dan Algoritma NN+SMOTE. AUC paling rendah adalah metode NN tanpa Resampling.

D. KESIMPULAN

Perusahaan asuransi adalah perusahaan yang bergerak di bidang jasa manajemen resiko. Dengan pengelolaan resiko yang baik akan membuat sebuah perusahaan asuransi berkembang maju.

Pengelolaan resiko berkaitan erat dengan pengenalan setiap resiko yang akan diterima, sehingga perlu adanya suatu metode yang baik dan relatif akurat untuk memprediksi baik buruknya suatu resiko.

Data produksi polis jika dibandingkan dengan klaim yang masuk akan sangat timpang sekali, karena semakin baik perusahaan asuransi dalam mengelola resiko maka semakin sedikit klaim yang terjadi.

Oleh karena itu jika membuat model prediksi dengan beberapa algoritma Data Mining, perlu hati-hati agar tidak terjebak ke dalam akurasi paradoks. Dimana variable dependant akan terus menunjukan kelas yang mayoritas setiap kali dilakukan prediksi dengan metode yang sudah dibuat.

Neural Network cukup handal dalam melakukan tugas pengenalaan pola dan prediksi di dalam bidang Data Mining, namun Neural Network memiliki kelemahan pada sensitifnya algoritma terhadap kelas yang tidak seimbang. Dari penelitian ini diketahui bahwa penggunaan Neural Network pada prediksi klaim memiliki akurasi yang tinggi, namun ternyata bila dilihat dari nilai AUC nya masuk ke dalam kelompok gagal karena di bawah ambang batas (0.6).

Penggunaan tehnik resampling pada data training cukup membantu dalam menaikan kinerja algoritma prediksi walaupun angka AUC nya masih di bawah nilai baik (di bawah 0.7). Namun secara umum ada peningkatan yang signifikan bila dibandingkan dengan tidak menggunakan tehnik resampling pada pengolahan data awal (pre-processing).

E. SARAN

Penelitian ini memberikan beberapa cara dalam menangani ketidak seimbangan kelas pada data untuk prediksi Klaim Asuransi Kecelakaan Diri Penumpang. Namun karena hasil yang didapat masih di bawah nilai baik, maka perlu adanya penelitian lebih lanjut mengenai hal ini, antara lain:

1. Penerapan feature selection pada data pre-processing, seperti Principle Component Analysis maupun SOM, atau tehnik featur selection lainnya.

2. Penggunaan algoritma lain yang lebih tahan terhadap kelas tidak seimbang, seperti beberapa varian dari algoritma Tree.
3. Penerapan penalized pada kelas target untuk menambah rentang probabilitas.

F. REFERENCES

[1] Afzal, W., & Torkar, R. (2008). Lessons from applying experimentation in software engineering prediction systems.

[2] Akdon dan Riduwan. 2005. Rumus dan Data dalam Aplikasi Statistika, Bandung: Alfabeta

[3] Andrea Dal Pozzolo (2011). Comparison of Data Mining Techniques for Insurance Claim Prediction. Universita degli Studi di Bologna.

[4] Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research.

[6] Chu-Siu Li (). Risk Clasification and Claim Prediction : An Empirical Analysis from Vehicle Damage Insurance in Taiwan.

[7] Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J. (2014). Analysis of sampling techniques for imbalanced data: An n=648 ADNI study. NeuroImage.

[8] Dr. Kasmir, SE. MM. (2013). Bank dan Lembaga Keuangan lainnya, PT Raja Grafindo Persada

[9] Freund, R. J., J, W. W., & L, M. D. (2003). Statistical Methods (Vol. 2). Academic Press.

[10] Harsih Rianto (2015) : Resampling Logistic Regression Untuk Penanganan Ketidakseimbangan Class Pada Prediksi Cacat Software. Nusa Mandiri. Jakarta

[11] Inna Kolyshkina and Marcel van Rooyen (2005). Text mining for insurance claim cost prediction. The Institute of Actuaries of Australia

[12] Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering.

[13] Janez Demsar (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7 (2006) 1–30

[14] Larose, D. T. (2005). Discovering Knowladge In Data: An Introduction to Data Mining. Discovering Knowledge in Data: An Introduction to Data Mining.

[15] Lijia Guo, Ph.D., ASA (2003). Applying Data Mining Techniques in Property~Casualty Insurance. University of Central Florida

[16] Maimon and Rokach (2010). Introduction to Knowledge Discovery and Data Mining

[17] Pelayo, L., & Dick, S. (2007). Applying novel resampling strategies to software defect prediction. Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS.

[18] Seymour Geisser (1993). Predictive Inference. Chapman & Hall, Inc

[19] Sofia Aftab (2013). Data Mining in Insurance Claims (DMICS) Two-way mining for extreme values. 978-1-4799-0615-4/13 ©2013 IEEE

[20] Thanathamathee, P., & Lursinsap, C. (2013). Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. Pattern Recognition Letters.

[21] Ripley & Venables (2012). Modern Applied Statistics with S. 4th Edition. Springer

[22] Vercellis, C. (2011). Business Intelligence: Data Mining and Optimization for Decision Making. Methods. John Wiley & Sons.

[23] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining Practical Mechine Learning Tools and Techniques Third Edition.

[24] Wu, X., & Kumar, V. (2010). The Top Ten Algorithms in Data Mining. Taylor & Francis Group.

[25] Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36.

[26] Yu, C. H. (2010). Resampling methods : Concepts, Applications, and Justification What is resampling? Types of resampling.

[27] Zhang, H., & Wang, Z. (2011). A normal distribution-based over-sampling approach to imbalanced data classification. In Artificial Intelligence and Lecture Notes in Bioinformatics.