METODE *DATA MINING* UNTUK PREDIKSI *CHURN* PELANGGAN

Yulianti
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pamulang
Jl. Surya Kencana No. 1, Pamulang, Tangerang Selatan-Indonesia
E-mail : yulianti.saifudin@gmail.com

Inti sari−Tingkat *churn* pelanggan di industri telekomunikasi cukup tinggi, sedangkan biaya untuk mendapatkan pelanggan baru jauh lebih mahal daripada mempertahankan pelanggan lama. Prediksi *churn* dapat digunakan oleh organisasi untuk mengidentifikasi pelanggan yang cenderung mengalami *churn*, sehingga dapat diambil tindakan untuk mempertahankannya. Untuk memprediksi pelanggan apakah akan mengalami churn atau tidak dapat dilakukan dengan menggunakan metode *data mining*. Pada penelitian ini dilakukan dengan menganalisa dataset pelanggan telekomunikasi menggunakan metode *data mining* untuk mencari metode terbaik yang dapat mengidentifikasi polanya. Hasil penelitian menunjukkan bahwa *Neural Network* (NN) merupakan model terbaik karena memiliki nilai akurasi tertinggi.

Kata Kunci : ***Data Mining, Churn Pelanggan, Prediksi***

Abstract−Level of customer churn in telecom industry is quite high, while the cost to acquire new customers is much more expensive than retaining existing customers. Churn prediction can be used by organizations to identify customers who tend to be churner, so can be taken an action to retain it. To predict whether the customer to be churn or not can be analysed by using data mining method. This research is done by analyzing telecommunication subscriber dataset using data mining method to find the best method that can identify the pattern. The results showed that Neural Network (NN) is the best model, because it has the highest accuracy value.

Keywords : Data Mining, Customer Churn, Prediction

1. PENDAHULUAN

Tiga puluh tahun yang lalu telepon seluler masih jarang digunakan, pelanggannya kurang dari 5 juta pelanggan di seluruh dunia [1]. Mereka cenderung menggunakan telepon seluler untuk dipasang di dalam mobil sebagai telepon mobil, yang banyak digunakan oleh para elit karena harga dan layanannya yang mahal, fasilitas yang disediakanpun hanya panggilan suara. Pada tahun 2015 diperkirakan mencapai lebih dari 7 milyar pengguna telepon seluler [2] [3], karena alat telekomunikasi dapat merangsang pertumbuhan ekonomi secara signifikan dan bahkan telah menjadi salah satu faktor keberhasilan pembangunan suatu bangsa. Dengan demikian telepon selular telah membawa masyarakat menuju kehidupan modern yang mengutamakan efisiensi dan kepraktisan.

Banyaknya operator seluler mendorong persaingan usaha yang sangat ketat. Pelanggan dapat memilih di antara beberapa penyedia layanan dan secara aktif menggunakan hak mereka beralih dari satu penyedia layanan ke yang lainnya. Terbukanya persaingan bebas diperusahaan jasa telekomunikasi merupakan salah satu tantangan serius yang harus dihadapi oleh industri telekomunikasi [4]. Kemudahan pelanggan untuk berpindah ke pesaing merupakan perhatian utama bagi bagian CRM (*Customer Relationship Management*) [5], karena untuk mendapatkan pelanggan baru biayanya lebih mahal lima sampai enam kali lipat daripada mempertahankan pelanggan yang sudah ada [6]. Hal ini juga telah menjadi isu penting dan merupakan salah satu tantangan utama perusahaan yang harus dihadapi di era global ini. Dalam pasar yang sangat kompetitif ini, pelanggan menuntut produk yang disesuaikan, dan layanan yang lebih baik dengan harga yang lebih murah, sementara penyedia layanan terus fokus pada akuisisi sebagai tujuan bisnis mereka.

Mengingat fakta bahwa industri telekomunikasi mengalami rata-rata tingkat *churn* tahunan 30-35 persen, dan biaya untuk merekrut pelanggan baru 5-10 kali lebih mahal daripada mempertahankan yang sudah ada, maka mempertahankan pelanggan menjadi lebih penting daripada mengakuisisi pelanggan [7]. *Churn* pelanggan mengacu pada hilangnya pelanggan secara periodik dalam suatu organisasi [8] [9]. *Churn* pelanggan merupakan penyebab kebocoran pendapatan terbesar dari perusahaan telekomunikasi [5]. Untuk mempertahankan pelanggan yang sudah ada, organisasi harus meningkatkan layanan pelanggan, memperbaiki kualitas produk, dan harus dapat mengetahui lebih awal pelanggan mana yang memiliki kemungkinan akan meninggalkan organisasi.

Prediksi *churn* dapat digunakan untuk mengidentifikasi *churners* lebih awal sebelum mereka berpindah, dan dapat membantu departemen CRM (*Customer Relationship Management*) untuk mempertahankan mereka, sehingga potensi kerugian perusahaan dapat dihindari [10]. Prediksi *churn* pelanggan merupakan strategi bisnis

yang penting bagi perusahaan. Dengan melakukan prediksi *churn* pelanggan, maka perusahaan dapat segera mengambil tindakan untuk mempertahankan pelanggan.

Prediksi dilakukan dengan cara menganalisa sekumpulan data yang besar untuk menemukan pola yang berguna dan kecenderungannya, proses ini disebut *data mining* [11]. *Data mining* adalah penelitian untuk mengumpulkan, membersihkan, mengolah, menganalisa, dan usaha mendapatkan wawasan yang berguna dari data [12]. *Data mining* telah menunjukkan kemajuan yang luar biasa dalam beberapa dekade terakhir, dan telah menjadi salah satu subbidang utama dalam ilmu komputer [13]. Telah banyak organisasi yang menggunakan *data mining* untuk mengelola hubungan dengan pelanggan, termasuk untuk mendapatkan pelanggan baru, meningkatkan pendapatan dari pelanggan yang ada, dan mempertahankan pelanggan yang memiliki nilai tinggi dan loyal [14].

Untuk melakukan prediksi menggunakan teknik *data mining* diperlukan data-data masa lalu yang telah dikumpulkan. Data-data konsumen banyak tersedia di dalam database perusahaan, bagaimana menggunakannya untuk memprediksi *churn* pelanggan merupakan tantangan bagi para peneliti [15]. Tetapi untuk mendapatkan dataset pelanggan yang sebenarnya merupakan masalah yang sulit bagi peneliti, karena dapat disalahgunakan [10]. Sehingga sebagian peneliti menggunakan dataset *churn* pelanggan yang telah disediakan untuk umum di internet.

Berdasarkan uraian di atas, pada penelitian ini akan diterapkan sejumlah algoritma klasifikasi dari metode *data mining* untuk mencari metode terbaik yang dapat digunakan untuk memprediksi *churn* pelanggan. Pada penelitian ini akan digunakan dataset yang tersedia di internet.

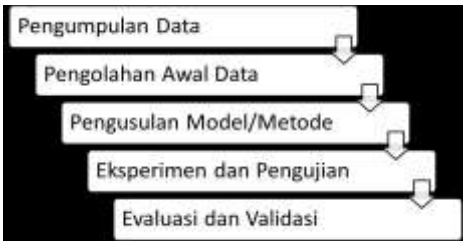
2. METODE PENELITIAN

Pada penelitian ini dilaksanakan untuk mencari pola dari pelanggan telekomunikasi yang mengalami *churn* dan tidak *churn* dengan menggunakan pendekatan kuantitatif. Metode kuantitatif berasal dari ilmu alam, di mana memiliki perhatian untuk memperoleh bagaimana sesuatu dikonstruksi, dibangun, atau bekerjanya [16]. Penelitian kuantitatif biasanya didorong oleh hipotesis yang dibuat dan diuji secara ketat untuk menunjukkan bahwa hipotesisnya tidak tepat. Sehingga tujuan utamanya adalah menyalahkan hipotesis, jika hipotesisnya tahan uji atau benar, maka disimpulkan hipotesis benar. Pengukuran merupakan dasar dari penelitian kuantitatif karena akan memberikan hubungan antara observasi dan formalisasi model, teori, dan hipotesis. Tujuan dari penelitian kuantitatif adalah mengembangkan model, teori, dan hipotesis yang berkaitan dengan fenomena alam.

Pada penelitian ini menggunakan metode eksperimen, yaitu mencakup investigasi terhadap hubungan sebab-akibat melalui pengujian yang terkontrol [17]. Pada penelitian eksperimental sering mendapatkan kendala pada ketidakcukupan akses terhadap

sampel, masalah pada etika, dan sebagainya. Eksperimen sering dilakukan dilakukan dalam pengembangan, evaluasi, dan pemecahan masalah proyek. Sebagian besar penelitian yang dilakukan di laboratorium menggunakan metode eksperimen [18]. Penelitian ini dilaksanakan dengan menganalisa menggunakan software RapidMiner terhadap dataset yang diambil dari internet, sehingga digunakan metode eksperimen.

Penelitian dilaksanakan untuk mendapatkan model prediksi terhadap pelanggan telekomunikasi apakah cenderung mengalami *churn* atau tidak. Penelitian ini dilakukan mengikuti tahapan pada Gambar 1, karena agar penelitian dapat diakui/diterima harus mengikuti aturan yang diakui [17].



Gambar 1 Tahapan Penelitian

Pengumpulan Data

Beberapa dataset tentang *churn* pelanggan industri telekomunikasi telah diunggah dan tersedia secara umum di internet. Sistem intelijen dan model matematis sebagai pendukung pengambilan keputusan dapat memberikan hasil yang akurat dan efektif hanya jika data yang digunakan dapat diandalkan [19]. Dataset yang digunakan pada penelitian ini adalah data sekunder, yaitu dataset *churn* pelanggan industri telekomunikasi yang diunduh dari https://bigml.com/dashboard/source/55c69eca200d5a25a0005180.
Dataset yang telah diunduh terdapat 3333 record dan 20 atribut. Spesifikasi dan atribut dataset *churn* pelanggan industri telekomunikasi yang diperoleh ditunjukkan pada Tabel 1.

Pengolahan Awal Data

Data *churn* pelanggan industri telekomunikasi yang telah dikumpulkan diolah untuk membuang data yang tidak relevan, dan data dengan atribut yang hilang. Pengolahan juga dilakukan dengan mengkonversi nilai-nilai redundan (berlebihan), atau mengelompokkan nilai yang terlalu beragam menjadi kelompok yang lebih kecil agar mempermudah pembentukan model. Pada penelitian ini semua data dan atribut yang telah diunduh digunakan dalam analisa *data mining*.

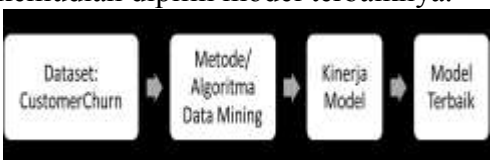
Tabel 1 Spesifikasi dan Atribut Dataset *Churn* Pelanggan

No.	Atribut	Keterangan	Contoh
1	State	Kode 51 negara bagian (*district*) Columbia	KS
2	Account Length	Lama akun aktif	128
3	Area Code	Kode area	415
4	Int'l Plan	Rencana/mengaktifkan panggilan internasional; 1=Ya 0=Tidak	0
5	VMail Plan	Rencana/mengaktifkan pesan suara (voice mail); 1=Ya 0=Tidak	1
6	VMail Message	Pesan suara	25
7	Day Mins	Lama panggilan siang hari (menit)	265,1
8	Day Calls	Jumlah panggilan siang hari	110
9	Day Charge	Biaya panggilan siang hari	45,07
10	Eve Mins	Lama panggilan sore hari (menit)	197,4
11	Eve Calls	Jumlah panggilan sore hari	99
12	Eve Charge	Biaya panggilan sore hari	16,78
13	Night Mins	Lama panggilan malam hari (menit)	244,7
14	Night Calls	Jumlah panggilan malam hari	91
15	Night Charge	Biaya panggilan malam hari	11,01
16	Intl Mins	Lama panggilan Internasional (menit)	10
17	Intl Calls	Jumlah panggilan internasional	3
18	Intl Charge	Biaya panggilan internasional	2,7
19	CustServ Calls	Jumlah panggilan ke layanan pelanggan (*customer service*)	1
20	Churn	Status churn (0=tidak/1=ya)	0

Pengusulan Model/Metode

Model prediksi *churn* pelanggan pada penelitian ini diusulkan menggunakan beberapa metode data mining, yaitu 10 algoritma klasifikasi. Untuk mengetahui model mana yang memiliki kinerja terbaik, maka akan diukur berdasarkan akurasinya, karena secara umum kinerja pengklasifikasi dievaluasi menggunakan keseluruhan akurasi model pada pengujian dataset [20].

Usulan model/metode ditunjukkan pada Gambar 2. Dataset yang telah diperoleh akan diterapkan pada metode/algoritma *data mining*. Kinerja dari setiap model diukur, kemudian dipilih model terbaiknya.



Gambar 1 Model/metode Usulan

Eksperimen dan Pengujian

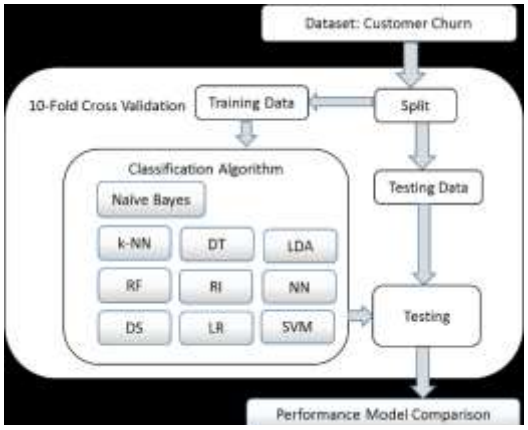
Eksperimen pada penelitian ini dilakukan menggunakan laptop untuk mengukur kinerja model yang diusulkan. Spesifikasi perangkat keras yang digunakan berupa sebuah laptop ACER menggunakan prosesor Intel (R) Pentium (R) P6200 @ 2.13GHz 64 bit, memori (RAM) 3,00 GB. Sedangkan perangkat lunak yang digunakan adalah Windows 10 Enterprise 64 bit sebagai sistem operasi, RapidMiner Studio 8.0.1 64 bit Free Edition untuk menerapkan dan menganalisa model yang diusulkan.

Evaluasi dan Validasi

Evaluasi dilakukan dengan cara mengamati hasil prediksi menggunakan algoritma data mining yang diusulkan. Sedangkan validasi untuk memastikan bahwa akan memberikan hasil yang sama ketika dilakukan secara mandiri. Untuk mengetahui metode/algoritma yang terbaik, maka dilakukan pengukuran dan membandingkan akurasi yang dihasilkan. Ukuran kinerja model dapat ditampilkan menggunakan *confusion matrix*. Nilai pada *Confusion matrix* diperoleh dari hasil validasi menggunakan *cross validation*.

Cross validaton merupakan metode statistik yang digunakan untuk mengevaluasi dan membandingkan *learning algorithms* (algoritma pembelajar) dengan cara membagi data menjadi dua bagian, satu bagian untuk belajar atau data latih, dan bagian yang lain untuk memvalidasi model [21]. Pada *cross validation* setiap data harus memiliki kesempatan tervalidasi, maka kumpulan pelatihan dan validasi dibuat *crossover*.

Teknik umum untuk mengukur kinerja pengklasifikasi adalah *K-fold cross validation*, dilakukan dengan cara menggunakan kembali dataset yang sama untuk menghasilkan k perpecahan dari dataset menjadi *non-overlapping* dengan proporsi pelatihan (k-1)/k dan 1/k untuk pengujian [22].



Gambar 2 Kerangka Kerja Penelitian

Pada penelitian ini digunakan teknik validasi *10-fold cross validation* untuk menguji model yang diusulkan, sehingga model dapat langsung diuji sebanyak 10 kali pengujian. Kerangka kerja penelitian *churn* pelanggan ini ditunjukkan pada Gambar 2.

Dataset yang diperoleh (*Customer Churn*) dipecah (*split*) menjadi data latih (*training data*) dan data uji (*testing data*) berdasarkan algoritma *10-fold cross validation*. Algoritma klasifikasi dilatih menggunakan data latih, kemudian diuji menggunakan data uji. Hasil pengujian ditampilkan menggunakan *confusion matrix*, kemudian digunakan untuk menghitung kinerja algoritma/model. Setelah semua algoritma/model diuji, selanjutnya dilakukan perbandingan kinerja untuk mendapatkan model terbaik.

3. HASIL DAN PEMBAHASAN

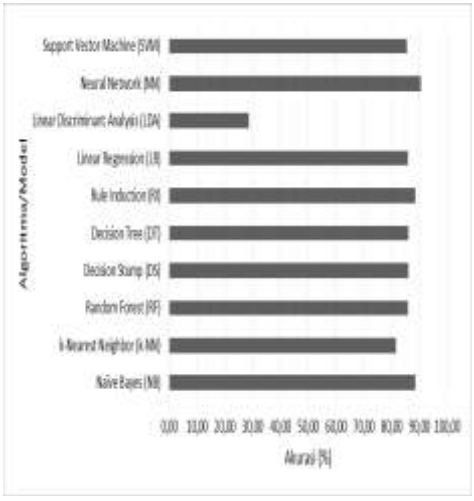
Hasil Pengukuran

Dataset yang diperoleh diterapkan pada model yang diusulkan menggunakan software RapidMiner Studio 8.0.1 64 bit Free Edition. Dari eksperimen yang dilakukan diperoleh kinerja (ukuran akurasi) seperti ditunjukkan pada Tabel 2.

Untuk mempermudah perbandingan hasil pengukuran kinerja model/algortima, maka divisualisasi menggunakan grafik tipe clustered chart seperti pada Gambar 3.

Tabel 2 Hasil Pengukuran Akurasi Algoritma/Model

No	Algoritma/model	Akurasi
1	Naïve Bayes (NB)	88,51%
2	k-Nearest Neighbor (k-NN)	81,61%
3	Random Forest (RF)	85,93%
4	Decision Stump (DS)	86,05%
5	Decision Tree (DT)	86,05%
6	Rule Induction (RI)	88,63%
7	Linear Regression (LR)	85,93%
8	Linear Discriminant Analysis (LDA)	28,67%
9	Neural Network (NN)	90,55%
10	Support Vector Machine (SVM)	85,48%



Gambar 3 Visualisasi Akurasi Algoritma/Model

Pembahasan

Berdasarkan hasil pengukuran menunjukkan bahwa *Neural Network* (NN) merupakan model terbaik dengan akurasi 90,55%. Sedangkan algoritma *Linear Discriminant Analysis* (LDA) tidak dapat digunakan untuk memprediksi *churn* pelanggan karena nilainya hanya 28,67%.

4. KESIMPULAN

Dari hasil penelitian menunjukkan bahwa *Neural Network* (NN) merupakan model terbaik karena memiliki nilai akurasi tertinggi. Sedangkan algoritma *Linear Discriminant Analysis* (LDA) gagal mengklasifikasikan karena memiliki akurasi kurang dari 60%.

DAFTAR PUSTAKA

[1] J. Cullen, M. Wahlqvist and G. Gómez, End-to-End Quality of Service over Cellular Networks. Data Services Performance and Optimization in 2G/3G, G. Gómez and R. Sanchez, Eds., West Sussex: John Wiley & Sons Ltd, 2005.

[2] B. Sanou, "ICT Facts & Figures," International Telecommunication Union, Geneva, 2015.

[3] R. K. Nistanto, "Tekno: 2015, Pengguna "Mobile" Lampaui Jumlah Penduduk Dunia," 4 Juni 2014. [Online]. Available: http://tekno.kompas.com/read/2014/06/04/1025003/2015.pengguna.mobile.lampaui.jumlah.penduduk.dunia.

[4] B. Huang, M. T. Kechadi and B. Buckley, "Customer Churn Prediction in Telecommunications," *Expert Systems with Applications,* pp. 1414-1425, 2012.

[5] R. J. Jadhav and U. T. Pawar, "Churn Prediction in Telecommunication Using Data Mining Technology," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 2, no. 2, pp. 17-19, 2011.

[6] W. Verbeke, K. Dejaeger, D. Martens, J. Hur and B. Baesens, "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," *European Journal of Operational Research,* vol. 218, no. 1, pp. 211-229, 2012.

[7] J. Lu, "Predicting Customer Churn in the Telecommunications Industry - An Application of Survival Analysis Modeling Using SAS," in *Proceedings of the Twenty-Seventh Annual SAS® Users Group International Conference*, Orlando, 2002.

[8] A. Churi, M. Divekar, S. Dashpute and P. Kamble, "Analysis of Customer Churn in Mobile Industry using Data Mining," *International Journal of Emerging Technology and Advanced Engineering,* vol. 5, no. 3, pp. 225-230, 2015.

[9] X. Yu, S. Guo, J. Guo and X. Huang, "An Extended Support Vector Machine Forecasting Framework for Customer Churn in E-Commerce," *Expert Systems with Applications,* vol. 38, no. 3, pp. 1425-1430, 2010.

[10] V. Umayaparvathi and K. Iyakutti, "Applications of Data Mining Techniques in Telecom Churn Prediction," *International Journal of Computer Applications,* vol. 42, no. 20, pp. 5-9, 2012.

[11] D. T. Larose and C. D. Larose, Data Mining and Predictive Analytics, 2nd ed., New Jersey: John Wiley & Sons, Inc., 2015.

[12] C. C. Aggarwal, Data Mining: The Textbook, Switzerland: Springer International Publishing, 2015.

[13] J. Stefanowski, "Dealing with Data Difficulty Factors While Learning from Imbalanced Data," in *Challenges in Computational Statistics and Data Mining*, Switzerland, Springer International Publishing, 2016, pp. 333-363.

[14] M. Madan, M. Dave and V. K. Nijhawan, "A Review on: Data Mining for Telecom Customer Churn Management," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 5, no. 9, pp. 813-817, September 2015.

[15] Z.-Y. Chen, Z.-P. Fan and M. Sun, "A Hierarchical Multiple Kernel Support Vector Machine for Customer Churn Prediction Using Longitudinal Behavioral Data," *European Journal of Operational Research,* vol. 223, no. 2, pp. 461-472, 2012.

[16] M. Berndtsson, J. Hansson, B. Olsson and B. Lundell, Thesis Projects: A Guide for Students in Computer Science and Information Systems, 2nd ed., London: Springer-Verlag, 2008.

[17] C. W. Dawson, Projects in Computing and Information Systems A Student's Guide, 2nd ed., Great Britain: Pearson Education, 2009.

[18] A. B. Badiru, C. F. Rusnock and V. V. Valencia, Project Management For Research: A Guide for Graduate Students, Boca Raton, Florida: CRC Press, 2016.

[19] C. Vercellis, Business Intelligence-Data Mining and Optimization for Decision Making, West Sussex: John Wiley & Sons, 2009.

[20] H. Zhang and Z. Wang, "A Normal Distribution-Based Over-Sampling Approach to Imbalanced Data Classification," in *Advanced Data Mining and Applications - 7th International Conference*, Beijing, 2011.

[21] P. Refaeilzadeh, L. Tang and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*, Arizona, Springer US, 2009, pp. 532-538.

[22] K. B. Korb and A. E. Nicholson, Bayesian Artificial Intelligence, 2nd ed., Florida: CRC Press, 2011.