ISSN :2085-6989

CHURN PREDICTION PELANGGAN MENGGUNAKAN CRISP-DM (Studi Kasus Pelanggan TelkomFlexi Bandung)

Customer Churn Prediction Using CRISP – DM (Case Study : Customer of TelkomFlexi Bandung)

Yance Sonatha
Jurusan Teknologi Informasi Politeknik Negeri Padang Kampus UNAND Limau Manis Padang 25163 Telp 0751-72590 Fax 0751-72576
E-mail : ys_line@yahoo.com , sonatha.yance@gmail.com

ABSTRACT

Nowadays, the need to stay ahead of the competencies is one of the company's focus. In an effort to stay afloat in the competitive conditions, companies can perform a variety of information technology development. One of the information technologies that are currently being intensively developed is the use of Data Mining. Data Mining provides benefits to process data into useful information for a company. The main pont of the research on this journal is to establish data mining models to identify TelkomFlexi postpaid customers who have the possibility to move to another provider (churn analysis).Churn analysis is done so that the company can choose the most appropriate strategy in retaining customers. Type of data mining is used to churn analysis is classification.

***Keywords** : Data Mining, CRISP DM, Churn Prediction, TelkomFlexi*

PENDAHULUAN

Data mining, atau juga dikenal dengan nama lainnya seperti *Knowledge Discoverying Databases* (KDD), *Knowledge extraction*, analisa pola data, arkeologi data, *data dredging*, *information harvesting*, *business intelligent*, dsb, adalah sebuah bidang ilmu yang berupaya menemukan pola, kaidah, aturan, dan informasi berharga yang menarik dan belum diketahui sebelumnya dari sekumpulan besar data. Kemunculan ilmu ini dilatarbelakangi oleh munculnya tumpukan data di berbagai bidang kehidupan. Seringkali sebuah organisasi atau kelompok kerja tertentu banyak melakukan kegiatan pengumpulan data, administrasi maupun perhitungan-perhitungan yang menghasilkan data dalam jumlah besar. Data ini dapat dimanfaatkan sebagai sumber informasi dengan memanfaatkan data mining.

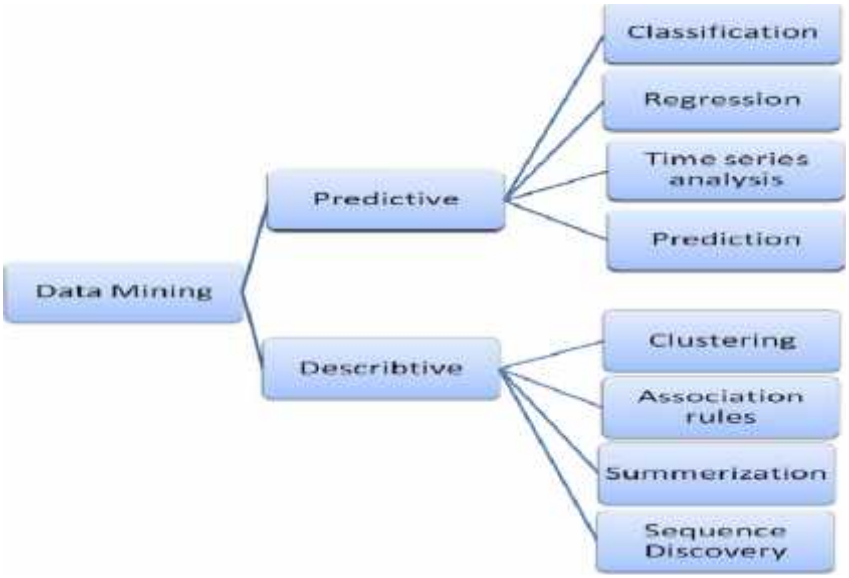
Berdasarkan teori dan prosedur *Cross-Industy Standard Process Data Mining* (CRISP DM), enam fase pengembangan *data mining*. Fase tersebut antara lain : *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluating*, dan *Deployment*.

LANDASAN TEORI

Data Mining

Error! Reference source not found. dibawah ini, menunjukkan fungsi-fungsi dari data mining.

ISSN :2085-6989



Gambar 1 Fungsi Data Mining

1) Klasifikasi, adalah fungsi pembelajaran yang memetakan (mengklasifikasi) sebuah unsur/item data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan.
2) Regresi, adalah fungsi pembelajaran yang memetakan sebuah unsur ke sebuah variabel prediksi bernilai nyata.
3) *Time series analysis*, mencari urutan kesamaan, pola, periode dan deviasi.
4) *Prediction*, memprediksi nilai-nilai yang mungkin terjadi dari data yang hilang atau distribusi nilai dari atribut tertentu dalam kumpulan objek.
5) *Clustering*/ Pengelompokan, merupakan tugas deskripsi yang banyak digunakan dalam mengidentifikasi sebuah himpunan terbatas pada kategori/ *cluster* untuk mendeskripsikan data yang ditelaah. Kategori-kategori ini dapat bersifat eksklusif dan ekshaustif mutual, atau mengandung representasi yang lebih kaya seperti kategori yang hirarkhis atau saling menumpu (*overlapping*).
6) *Association rules*, menemukan hubungan dan korelasi antar berbagai *item*.
7) Pencarian pola sekuensial, menganalisa sekumpulan record pada suatu periode waktu, misalnya untuk menganalisa trend.

Classification

Classification bertujuan untuk memetakan (mengklasifikasi) sebuah unsur/item data ke dalam salah satu dari beberapa kelas yang sudah didefinisikan. Ada dua hal yang harus dipertimbangkan dalam pelaksanaan *classification.* Hal yang pertama adalah model *classifier* yang akan digunakan, dan yang kedua adalah cara pengujian *classifier* tersebut. Pengujian dilakukan untuk menilai kinerja dari suatu *classifier*.

Classifier adalah model atau algoritma yang digunakan untuk menklasifikasi data. Beberapa *classifier* yang akan digunakan pada proyek ini antara lain :

KNN (K-Nearest Neighbor*)*

Metode KNN merupakan salah satu pendekatan yang digunakan dalam pengklasifikasian secara mudah dan efisien. Konsep dasar algoritma KNN adalah mencari jarak terdekat antara data yang dievaluasi dengan sejumlah K tetangga (*neighbor*) terdekatnya dalam data uji.

Support Vector Machine (SVM)

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali diperkenalkan pada tahun 1992 di "Annual Workshop on Computational Learning Theory". Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti *margin hyperplane* (Duda & Hart tahun 1973, Cover tahun 1965, Vapnik 1964, dsb.), dan Kernel yang diperkenalkan oleh Aronszajn pada tahun 1950. SVM berusaha menemukan *hyperplane* yang terbaik pada suatu *input space*. Prinsip dasar SVM adalah *linear classifier*, dan selanjutnya dikembangkan agar dapat bekerja pada problem *non-linear*. Dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi.

Naïve Bayes

Teori Bayesian diadopsi dari nama penemunya yaitu Thomas Bayes . Teori ini diperkenalkan pada sekitar tahun 1950, dan semenjak itu sering digunakan pada studi-studi ilmu statistika yang berbasis pada teorema atau aturan Bayes.

Teori Bayesian adalah sebuah teori kondisi probabilitas yang memperhitungkan probabilitas suatu kejadian (hipotesis) bergantung pada kejadian lain (bukti). Pada dasarnya, teorema tersebut mengatakan bahwa kejadian di masa depan dapat diprediksi dengan syarat kejadian sebelumnya telah terjadi [Andriansyah, 2005]. Model Naïve Bayes banyak digunakan untuk *clustering* dan *classification*.

Alternating Decision Tree (AD Tree)

Decision tree adalah diagram alir (*flowchart*)berbentuk seperti cabang pohon, dimana setiap titik percabangan menunjukkan sebuah *test* pada atribut, tiap cabang menunjukkan hasil dari test dan *leaf-node* menunjukkan *class* atau *class distribution* [Sunjana, 2010].

Alternating decision tree(ADTree) diperkenalkan pertama kali oleh Yoaf Freund dan Llew Mason. Metode ini menyediakan mekanisme yang mengkombinasikan hipotesis secara umum yang ditujukan untuk meningkatkan sebuah representasi yang tunggal (Holmes,2008).

PERMASALAHAN

Proyek pengembangan data mining kali ini dilakukan dengan menggunakan studi kasus pada perusahaan Telkom, dengan produk andalannya Telkom Flexi. Hasil akhir dari proyek ini antara lain *clustering* pelanggan Telkom Flexi dan prediksi pelanggan yang memiliki kemungkinan berpindah ke provider lain (*churn*). Diharapkan berdasarkan hasil dari proyek ini, perusahaan dapat mengidentifikasi pelanggan yang berpotensi berpindah ke provider lain, sehingga perusahaan dapat membangun strategi baru yang lebih komprehensif untuk mempertahankan pelanggan.

METODOLOGI

Ada berbagai pendekatan untuk mengukur kinerja *classifier*. Secara sederhana pengukuran dapat dilakukan dengan menghitung proporsi *predicted examples*yang secara benar diprediksi oleh *classifier* dengan jumlah data seluruhnya.

Pada proyek ini hasil dari *classifier* akan diuji berdasarkan pilihan yang ada pada"test options box" di aplikasi Weka 3.6. Terdapat 4 jenis pengujian yang akan digunakan, yaitu:

a) *Use Training Set*
   Classifier dievaluasi dalam hal seberapa baik *classifier* tersebut dalam memprediksi kelas dari data *training*.
b) *Supplied Test Set*
   Classifier dievaluasi dalam hal seberapa baik *classifier* tersebut dalam memprediksi kelas dari sekumpulan *instances* yang dimuat dari sebuah *file*.
c) Cross Validation
   Classifier di evaluasi dengan dengan *cross validation*, mengunakan jumlah

dari *folds* yang dimasukkan dalam *folds text field*.Karakteristik *cross validation* :
- Sumber data dipisahkan menjadi beberapa *folder* untuk *cross validation*.
- Seleksi fitur-fitur dilakukan pada masing-masing folder training data secara terpisah.
- Beragam tree diperoleh dari beberapa kasus.
- Evaluasi klasifikasi dilakukan berdasarkan semua hasil yang diperoleh.

d) Percentage Split

Classifier dievaluasi dalam hal seberapa baik *classifier* tersebut dalam memprediksi persentase pasti dari data yang digunakan untuk *testing*. Jumlah dari data yang digunakan tergantung pada nilai yang dimasukkan dalam *field %* pada Weka 3.6.

Percentage split merupakan salah satu pilihan dalam test option. Jumlah data tergantung pada nilai % yang dimasukkan kolom (Kirkby and Frank, 2004).

Adapun parameter pengukuran yang digunakan adalah :
- *Accuracy* adalah persentase dari total data yang benar diidentifikasi.
- *Precision* adalah perbandingan jumlah data relevan yang diambil dengan jumlah seluruh dokumen yang diambil oleh sistem (Mandala, 2006).
- *Recall* merupakan salah satu pengukuran untuk permasalahan dalam klasifikasi.*Recall* adalah perbandingan antara jumlah data relevan yang diambil dengan jumlah data relevan yang berada di koleksi dokumen (database).
- *F Measures* merupakan sebuah parameter pengukuran yang membandingkan secara seimbang antara *precision* dan *Recall*.

PEMBAHASAN

Pengumpulan Data

Data untuk analisis *churn* diambil dari sistem operasional yang ada di PT. Telkom. Pada proyek ini, data *customer* yang digunakan sebanyak 16.696 record. Ada empat jenis data yang diambil, yaitu: data demografis, data traffic, data pembayaran dan data revenue.

Persiapan Data

1. Data Cleaning

Tahap *data cleaning* tidak dilakukan dalam proyek ini dikarenakan perusahaan telah melakukan pembersihan data untuk masing-masing database operasional, sehingga data yang diperoleh untuk proyek sudah merupakan data yang *correct* (tidak memiliki *error* dan *missing value*), sehingga data tersebut sudah dapat dikatakan sebagai data yang bersih, komplit dan terverifikasi.

Data Construction

Tahap data construction tidak dilakukan, karena tidak ada atribut yang didapatkan dari hasil pengolahan atribut lain. Keseluruhan atribut merupakan atribut murni dari data, yang didapatkan langsung dari database.

Integrate Data

Integrasi data dilakukan dengan menggabungkan data-data yang dipeoleh menjadi sebuah dataset pelanggan flexi pasca bayar. Data-data yang diintegrasikan adalah data demografis, data *traffic*, data pembayaran, data *revenue.*

Pada hasil integrasi, dataset ini memiliki kondisi *imbalance class.* Artinya terdapat ketidak seimbangan porsi data latih antara sebuah kelas dengan kelas yang lain. Dalam hal ini adalah porsi data latih antara kelas churn dan active, dimana data churn jauh lebih sedikit daripada data active. Permasalahan ini terjadi dikarenakan fakta/kejadian dari kelas tersebut memang jarang terjadi. *Imbalance class* menjadi suatu permasalahan penting karena mesin *learning* cenderung untuk menghasilkan

akurasi prediksi yang baik pada kelas data yang banyak (mayoritas) tetapi menghasilkan akurasi prediksi yang buruk pada kelas yang sedikit (minoritas). Hal tersebut dikarenakan kondisi data yang jarang menyebabkan algoritma *data mining* murni cenderung akan mengabaikannya sehingga rule / pola-pola yang terkandung di kelas minor tidak terekstrak dengan baik.

Metode yang digunakan untuk mengatasi permasalahn *imbalance class* yang terjadi adalah databoostIM yang memadukan *boosting* dengan *data generation, data generation* adalah membuat data sintetis untuk ditambahkan ke data asli (data sintetik yang dibuat sama dengan data asli). Sehingga akurasi prediksi terhadap kelas mayor dan kelas minor dapat diperbaiki.

Format data

Format data hanya dilakukan pada penamaan atribut tanpa mengubah makna aslinya, seperti atribut A1 untuk kategori cluster yang didefinisikan oleh *carrier*, A2 untuk jenis layanan yang dipakai pelanggan, A3 untuk jenis kelamin, dan seterusnya.

Pembuatan Model dan Evaluasi

Data aktual yang digunakan pada aplikasi ini sebanyak 16.696 *record*. Atribut-atribut yang dipakai untuk membangun model ada 21 atribut. Data ini memiliki komposisi kelas aktif sebanyak 16490 (98,77 %) dan kelas *churn* sebanyak 206 (1,23 %).

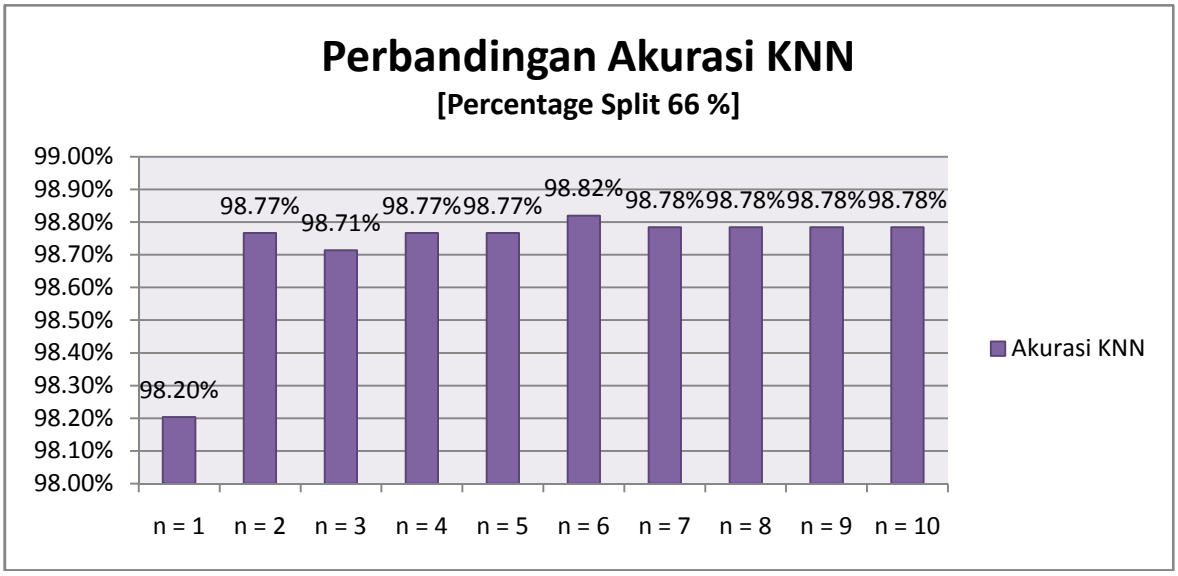
Klasifikasi *churn prediction* dilakukan dengan menggunakan empat *classifier*, yaitu Naïve Bayes, ADTree, KNN dan SVM. Pengujian dilakukan dengan membagi dataset menjadi data *training* dan data *testing*. Data *training* digunakan untuk membangun model, sedangkan data *testing* digunakan untuk menghitung akurasi model yang dibangun oleh *classifier*.

Metode testing *percentage split* 66% akan membagi dataset menjadi data training sebanyak 66% dan data testing sebanyak 33%. Komposisi ini dipilih berdasarkan pada *Data Mining : Practical Machine Learning Tools and Techniques* [Ian H. Written and Eibe Frank, 2005] yang menyebutkan bahwa data *training* yang digunakan sebaiknya berjumlah lebih dari setengah bagian dataset, yaitu dua per tiga bagian dari jumlah dataset, sedangkan satu per tiga bagian sisanya digunakan untuk data *testing*.

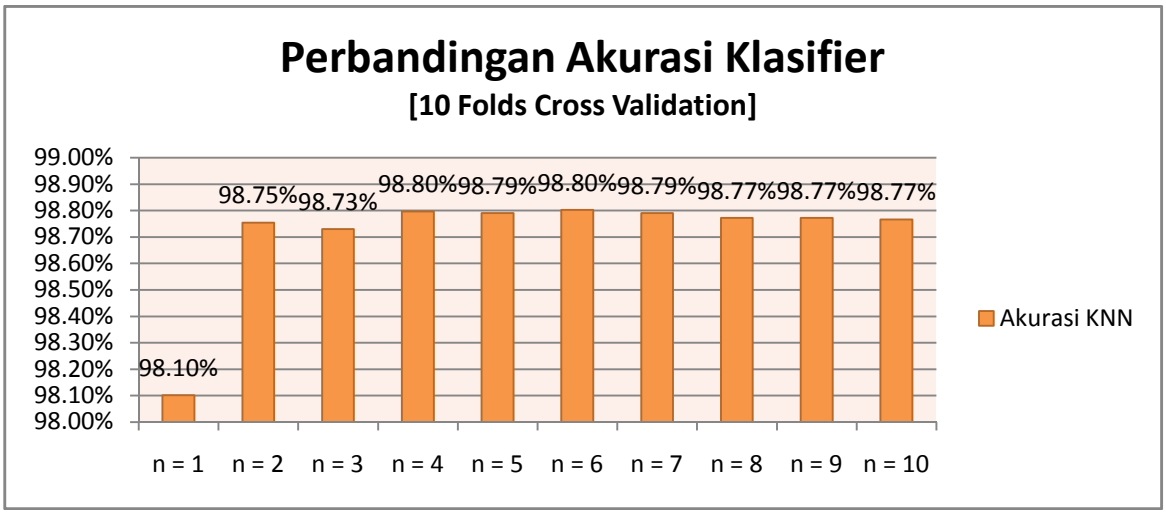
Metode *testing n folds cross validation* akan membagi dataset menjadi n bagian yang sama. Dalam pengujian ini dipilih n = 10. Nilai n ini dipilih berdasarkan pada *Data Mining : Practical Machine Learning Tools and Techniques* [Ian H. Written and Eibe Frank, 2005) yang menyebutkan bahwa 10 adalah nilai yang tepat untuk memperoleh estimasi kesalahan yang paling baik. Proses yang dilakukan adalah membagi data secara random menjadi 10 bagian yang sama, dimana 9/10 bagian akan digunakan sebagai data *testing*, sedangkan 1/10 digunakan sebagai data *training*. Iterasi proses *learning* dilakukan sebanyak 10 kali pada data *training* yang berbeda dengan mengubah formasi data training dan data testing. Nilai akhir akurasi dan estimasi kesalahan diperoleh dengan merata-ratakan keseluruhan nilai akurasi dan estimasi kesalahan yang diperoleh dari 10 kali iterasi yang dilakukan.

1. Perbandingan Akurasi *Classifier* KNN

Classifier KNN yang akan digunakan memiliki nilai n = 6 . nilai ini dipilih berdasarkan percobaan yang dilakukan dengan menggunakan nilai n = 1 sampai dengan n = 10, dimana nilai n = 6 memberikan akurasi yang paling baik dibandingkan dengan yang nilai n lainnya. Akurasi KNN dengan nilai n = 6 mencapai 98,82 %, untuk metode pengujian *percentage split 66 %* (Gambar 2) dan 98,80 % untuk metode pengujian *10 Folds Cross Validation* (Gambar 3).



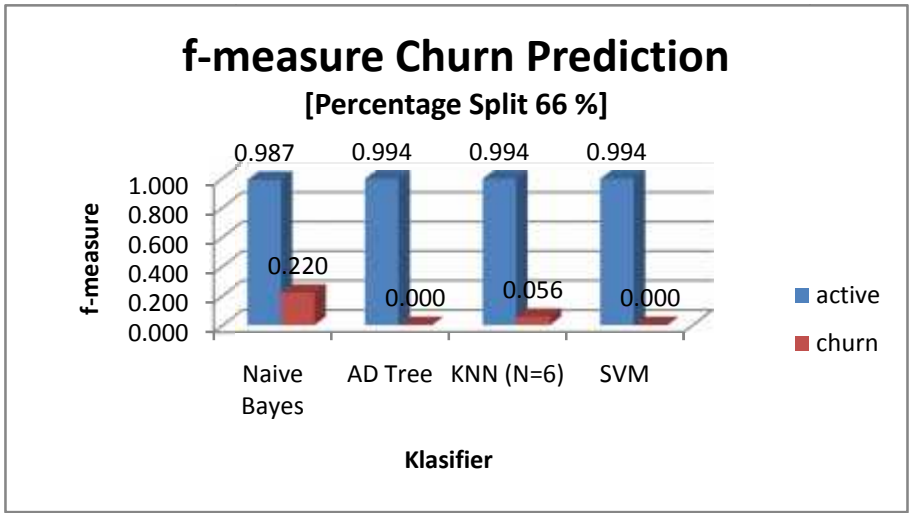
Gambar 2 Perbandingan Nilai Akurasi KNN Pada Metode Pengujian Percentage Split 66 %



Gambar 3 Perbandingan Nilai Akurasi KNN Pada Metode Pengujian 10 Folds Cross Validation

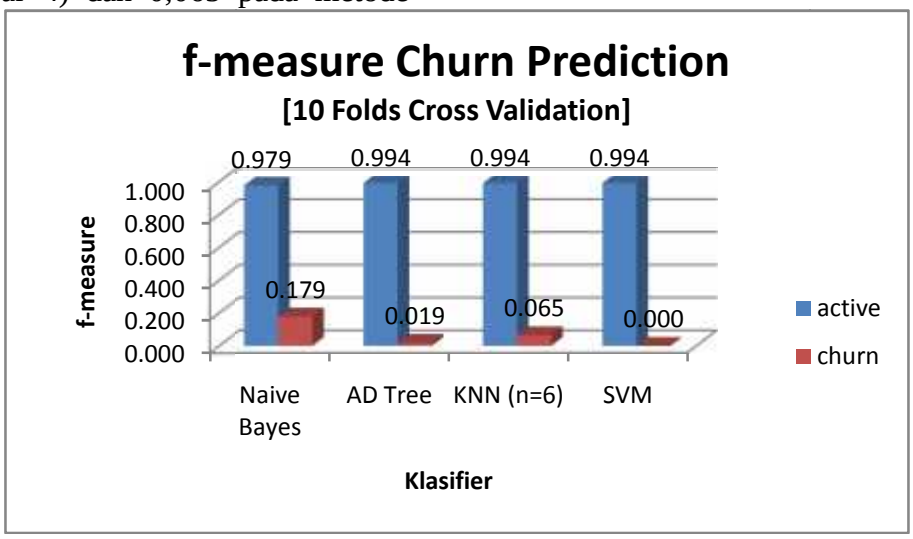
2. Analisa f-measure Naïve Bayes, ADTree, KNN, dan SVM untuk Churn Prediction

Pengukuran performansi *classifier* untuk *churn prediction* dinilai dengan *f-measure*. Berdasarkan hasil pengujian, dapat dikatakan bahwa keempat *classifier* memiliki performansi yang cukup baik dalam memprediksi pelanggan yang *active*. Hal ini ditunjukkan pada Gambar 4 untuk metode pengujian *percentage split 66 %* dan Gambar 5 untuk metode pengujian 10 *cross folds validation*, dimana nilai *f-measure* keempat *classifier* mencapai berada di kisaran 0,99.



Gambar 4 F-Measure Classifier untuk Churn Prediction pada Metode Pengujian Percentage Split 66 %

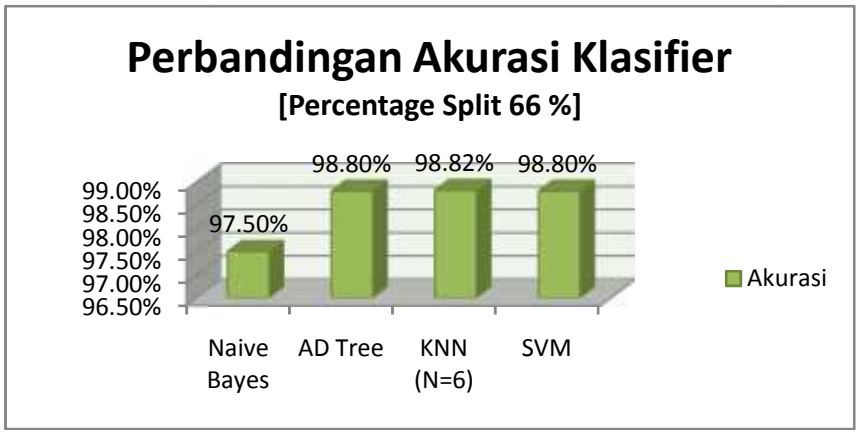
Berdasarkan hasil pengujian, Naïve Bayes memiliki nilai *f-measure* yang unggul dalam memprediksi pelanggan yang *churn* dibandingkan dengan tiga *classifier* (ADTree, KNN, SVM) lainnya. *F-measure* yang dicapai oleh Naïve Bayes sebesar 0,220 pada metode pengujian *percentage split 66%*(Gambar 4) dan 0,179 pada metode pengujian 10 *cross folds validation* (Gambar). Sementara itu, walapun *f-measure* KNN cukup rendah, 0,056 pada metode pengujian *percentage split 66%*(Gambar 4) dan 0,065 pada metode pengujian 10 *cross folds validation* (Gambar), namun model yang dibangun KNN masih dapat memprediksi pelanggan yang *churn.* Sedangkan model yang dibangun oleh ADTree hampir tidak dapat melakukan *churn prediction*. Hal ini tampak pada nilai *f-measure* yang diperoleh hampir mendekati nol. Pada *classifier* SVM, bahkan model yang dibangun tak dapat melakukan *churn prediction* (nilai *f-measure* untuk kedua metode pengujian adalah nol).



Gambar 5 F- measure Classifier untuk Churn Prediction pada Metode Pengujian 10 Folds Cross Validation

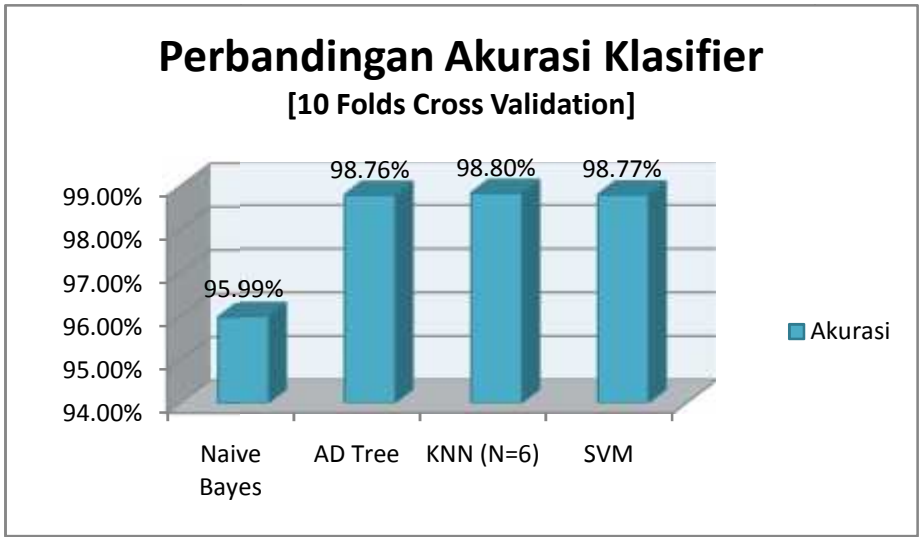
3. Analisa Akurasi Naïve Bayes, ADTree, KNN, dan SVM

Berdasarkan hasil metode pengujian *percentage split 66 %*, (Gambar), *classifier* ADTree, KNN, dan SVM memiliki nilai akurasi yang hampir sama yaitu berkisar pada 98,8%. Sedangkan akurasi yang lebih rendah dimiliki oleh Naïve Bayes, yaitu sebesar 97,50%. Hal yang hampir serupa juga diperoleh dari hasil metode *10 Folds Cross Validation* (Gambar), dimana *classifier* ADTree, KNN, dan SVM juga memiliki nilai akurasi yang hampir sama yaitu sebesar 98,8%. Sedangkan akurasi yang lebih rendah dimiliki oleh Naïve Bayes, yaitu sebesar 95,99%. Walaupun ada sedikit perbedaan nilai akurasi yang dicapai diantara keempat *classifier* yang digunakan, dapat dikatakan bahwa keempat *classifier* memiliki nilai akurasi yang cukup baik dalam melakukan klasifikasi pelanggan. Akan tetapi baiknya nilai akurasi ini dipengaruhi oleh dominasi keberhasilan *classifier* dalam memprediksi pelanggan yang aktif. Kondisi ini menutupi performansi *classifier* dalam memprediksi pelanggan yang *churn*, dimana kemampuan keempat *classifier* dalam memprediksi pelanggan yang *churn* tak sebaik kemampuan dalam memprediksi pelanggan yang aktif (berdasarkan Gambar dan Gambar).



Gambar 6 Akurasi Classifier Pada Metode Pengujian Percentage Split 66 %

Kelemahan *classifier* dalam melakukan *churn prediction* ini tidak lepas dari kondisi data aktual yang digunakan. Komposisi pelanggan yang aktif dan pelanggan yang *churn* tidak sebanding yaitu sebanyak 16490 (98,77 %) untuk kelas aktif dan 206 (1,23 %) untuk kelas *churn*. Hal inilah yang menyebabkan model yang dibangun oleh *classifier* akan condong merepresentasikan kelas *active.* Sehingga perbedaan nilai yang ditimbulkan atribut tertentu yang mungkin merepresentasikan kelas *churn* tidak dapat ditangkap oleh model karena mungkin saja nilai pembeda ini diabaikan oleh model yang didominasi sebagai kelas *active*.



Gambar 7 Akurasi Classifier Pada Metode Pengujian 10 Folds Cross Validation

KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan menggunakan empat *classifier*, yaitu Naïve Bayes, ADTree, KNN dan SVM, serta dengan menggunakan metode pengujian Percentage Split 66 % dan 98,80 % untuk metode pengujian 10 Folds Cross Validation, diperoleh kesimpulan sebagai berikut:

1) KNN dengan menggunakan metode pengujian percentage split 66% yaitu sebesar 98,82% menghasilkan hasil lebih baik dibandingkan dengan metode pengujian 10 Folds cross validation yaitu sebesar 98,80%.
2) Clasifier terbaik yang dapat digunakan untuk memprediksi kemungkinan *churn* pelanggan adalah Naïve Bayes, dengan *f-measure* paling baik apabila dibandingkan dengan metode ADTree, KNN dan SVM , yaitu sebesar 0,220 pada metode pengujian *percentage split 66%*(Gambar 3) dan 0,179 pada metode pengujian 10 *cross folds validation*.
3) Secara umum *classifier* Naïve Bayes, ADTree, KNN, dan SVM memiliki nilai akurasi cukup baik dalam melakukan klasifikasi pelanggan dengan nilai akurasi berkisar antara 97-98 %.
4) Kelemahan *classifier* dalam melakukan *churn prediction* tidak lepas dari kondisi data aktual yang digunakan. Komposisi pelanggan yang aktif dan pelanggan yang *churn* tidak sebanding yaitu sebanyak 16490 (98,77 %) untuk kelas aktif dan 206 (1,23 %) untuk kelas *churn*. Hal inilah yang menyebabkan model yang dibangun oleh *classifier* akan condong merepresentasikan kelas *active.*

REFERENSI

Andriansyah. 2005. "Metode Penyaringan Email yang Tidak Diinginkan Menggunakan Pendekatan Probabilitik". Seminar Nasional Aplikasi Teknologi Informasi 2005 pp. 19-23.

Holmes, G .et,al. 2008."Multiclass Alternating Decision Tree". Departemen of Computer Science University of Waikato. New Zealand.

Sunjana. 2010. "Aplikasi Mining Data Mahasiswa dengan Metode Klasifikasi Decision Tree". Seminar Nasional Aplikasi Teknologi Informasi. Yogyakarta.

Written, I. H. Dan Frank, E. 2005. "Data Mining : Practical Machine Learning Tools and Techniques 2nd edition".

San Francisco : Morgan Kaufmann Publisher.

Kirkby, R. dan Frank, E. 2004. "WEKA Explorer User Guide for Version 3-4-3". University of Waikato.

Mandala, R. 2006. "Evaluasi Efektifitas Metode Machine-Learning pada Search-Engine". Seminar Nasional Aplikasi Teknologi Informasi, pp. G11-G16.

Nugroho, A.M., Handoko, D., dan Witarti, B.A. 2005. "Analisa Informasi Dimensi Tinggi pada Bioinformatika Memakai Support Vector Machine". Proc. of National Conference on Information & Communication Technology (ICT) for Indonesia/e-Indonesia Initiatives-II, pp.427-435.

Trisedya dan Jais. 2009. "Klasifikasi Dokumen Menggunakan Algoritma Naïve Bayes dengan Penambahan Parameter Probabilitas Parent Category". Fakultas Ilmu Komputer Universitas Indonesia

Yahdin. 2008. "Aplikasi Pengambilan Keputusan pada Perencanaan Produksi Berdasarkan Teorema Bayes". Media Informatika, 6(1), pp. 25-38.