MEMPREDIKSI PEMBAYARAN BIAYA KURSUS SISWA ENTER (ENGLISH CENTER) DENGAN ALGORITMA K-NN

Sri Hadianti[1], Eka Rahmawati[2],Ilham Kurniawan[3], Windu Gata[4]
[1]Jurusan Ilmu Komputer, Fakultas Magister Ilmu Komputer, STMIK NUSA MANDIRI
Jalan Kramat Raya No.18, Jakarta 10420 Telp.(021) 31908575
Email: srihadianti07@gmail.com
[2] Jurusan Ilmu Komputer, Fakultas Magister Ilmu Komputer, STMIK NUSA MANDIRI
Jalan Kramat Raya No.18, Jakarta 10420 Telp.(021) 31908575
Email: ekarahawatiflash@gmail.com
[3] Jurusan Ilmu Komputer, Fakultas Magister Ilmu Komputer, STMIK NUSA MANDIRI
Jalan Kramat Raya No.18, Jakarta 10420 Telp.(021) 31908575
Email: lopioxo@gmail.com
[4]Jurusan Ilmu Komputer, Fakultas Magister Ilmu Komputer, STMIK NUSA MANDIRI
Jalan Kramat Raya No.18, Jakarta 10420 Telp.(021) 31908575
Email: Windu@nusamandiri.ac.id

ABSTRAK

ENTER (English Center) merupakan lembaga kursus bahasa inggris yang sangat peduli terhadap masyarakat dari semua kalangan yang berminat untuk belajar bahasa inggris. Pada lembaga kursus ENTER untuk meringankan siswa dalam pembayaran kursus, maka diberlakukan sistem pembayaran dengan metode cicilan, akan tetapi dengan metode tersebut banyak siswa yang membayar biaya kursus tidak tepat waktu sehingga menghambat pada biaya operasional perusahaan. Sehingga perlu adanya prediksi siswa yang akan melakukan pembayaran biaya kusus. Pada paper ini peneliti menggunkan algoritma K-NN yang diterapkan pada data siswa yang melakukan pembayaran biaya kursus di ENTER (Emglish Center). Hasil testing untuk mengukur performa algoritma ini menggunkan metode Cross Validation, Confussion Matrix dan kurva ROC dengan K5, K10, K15,dan K20 berturut-turut 84.69%, 83.93%, 83.36%, 83.36% dan menghasilkan akurasi dan nilai AUC 0.962, 0.937, 0.922, 0.916. Karena nilai AUC berada dalam rentang 0,9 sampai 1,0 maka metode tersebut masuk dalam kategori sangat BAIK.

Kata kunci: ENTER;CRISP-DM;K-NN;Cross Validation; Confusion Matrix; kurva ROC

PENDAHULUAN

Latar Belakang

Bahasa inggris merupakan bahasa internasional sehingga sangat penting untuk kita pelajari, apalagi di era Masyarakat Ekonomi ASEAN (MEA) sekarang yang dengan mudah nya orang luar masuk ke Indonesia, dan atau sebaliknya kita bisa dengan mudah masuk ke negara lain, akan tetapi kita butuh persiapan selain keahlian dalam bekerjadi butuhkan juga keahlian dalam bahasa inggris yang baik, sebab dengan bahasa inggris kita bisa berkomunikasi dengan orang dari negara lain.

Berdasarkan data dari survei "*English Proficiency Index*" (EF EPI) Indonesia berada di peringkat 32 dari 72 negara untuk kategori penguasaan Bahasa inggris pada skala Internasional. Indonesia memiliki nilai total 52,91 dalam mengukur kemampuan Bahasa Inggris negara-negara di dunia dan dianggap sebagai patokan internasional untuk kemampuan Bahasa Inggris tingkat dewasa. Hasil survei tahun 2016 menunjukan peringkat tiga besar negara dengan penguasaan Bahasa Inggris tertinggi di Asia adalah Singapura, Malaysia dan disusul Filipina. Di sisi lain, Indonesia meraih nilai yang lebih rendah dibandingkan beberapa negara tetangga di kawasan, termasuk Vietnam yang berada di posisi ke-31 yang tergolong "level menengah" (www.antaranews.com,2016). Dilihat dari data tersebut, maka kita bisa mengetahui bahwa di negara Indonesia penguasaan Bahasa Inggris masih berada di level menengah,dari data tersebut maka kita bisa

menyimpulkan bahwa sangat pentingnya keahlian bahasa inggris untuk masyarakat Indonesia, supaya bisa bersaing dengan negara tetangga yang lebih bagus penguasaan bahasa inggrisnya.

ENTER (English Center) merupakan lembaga kursus bahasa inggris yang sangat peduli terhadap semua kalangan yang berminat untuk belajar bahasa inggris, sehingga ENTER bisa membantu masyarakat untuk meningkatkan kemampuannya dalam bahasa inggris, ini terbukti dnegan dibukanya lebih dari 10 cabang di seluruh Indonesia.Pada lembaga Kursus ENTER untuk memudahkan masyarakat belajar bahasa inggris, maka di berlakukan sistem pembayaran dengan metode cicilan, sehingga masyarakat dikalangan menengah ke bawah tidak terlalu terbebani untuk bisa belajar, akan tetapi dengan metode tersebut menjadi suatu boomerang bagi lembaga, dikarenakan dengan banyaknya yang melakukan metode cicilan dan tidak melakukan pembayaran tepat waktu, sehingga menghambat pada operasional perusahaan (Kementrian Pendidikan, 2003)..

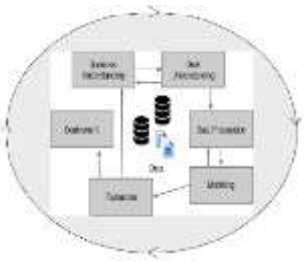
Berdasarkan latar belakang yang telah disebutkan diatas, didapatkan rumusan masalah yang dibahas dalam penelitian ini sebagai berikut:

3. Bagaimana menerapkan algoritma k-nn untuk memprediksi tipe siswa yang melakukan pembayaran kursus dengan cicilan?
4. Apakah algoritma k-nn dapat membantu memprediksi siswa yang akan melakukan pembayaran kursus dengan cicilan?

Tujuan dilakukannya penelitian ini adalah untuk mengetahui apakah pengolahan data prediksi siswa yang akan melakukan pembayaran kursus dengan cicilan menggunakan algoritma K-NN dapat membatu dalam menentukan strategi yang akan diambil oleh lembaga kursus bahas inggris ENTER berikutnya.

Untuk mempermudah penelitian maka dilakukan Data Mining. *Data Mining* (Witten, 2011) didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, *data mining* dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, *clustering* dan asosiasi (Larose, 2005). Proses dalam tahap *data mining* (Gambar 1.) terdiri dari tiga langkah Utama (Sumathi, 2006), yaitu :

a. *Data Preparation*
   Pada langkah ini, data dipilih, dibersihkan, dan dilakukan *preprocessed* mengikuti pedoman dan *knowledge* dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh.
b. Algoritma *data mining*
   Penggunaan algoritma *data mining* dilakukan pada langkah ini untuk menggali data yang terintegrasi untuk memudahkan identifikasi informasi bernilai.
5. Fase analisa data



Gambar 1. Siklus CRISP-DM (Chapman: 2000)

Keterangan siklus CRISP-DM adalah sebagai berikut:

a. Business Understanding
   Menentukan tujuan dan mendefinisikan masalah dari data mining
b. Data Understanding
   Mengumpulkan data awal dan identifikasi data kualitas
c. Data Preparation
   Menyiapkan data awal, kumpulan dan yang akan digunakan untuk keseluruhan fase berikutnya atau proses s eleksi data, pilih kasus dan variabel yang akan dianalisis sesuai dengan analisis yang akan dilakukan, lakukan perubahan pada variabel juka diperlukan, siapkan data awal hingga siap untuk perangkat pemodelam atau data transformation.

d. Modeling

Memilih dan mengaplikasikan model yang sesuai, kalibrasi aturan model untuk mengoptimalkan hasilm dapat menggunakan beberapa teknik yang sama untuk permasalahan yang sama, dapat kembali ke fase pengolahan data jika diperlukan untuk menjadikan data kedalam bentuk kebutuhan tertentu.

e. Evaluation
Evaluasi dan hasil agar selaras dengan tujuan bisnis

f. Deployment
Implementasi (penyebaran) dari data mining

Dalam teknik data mining perlu diadakan klasifikasi data. Klasifikasi data adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Han, 2006). Algoritma klasifikasi yang banyak digunakan secara luas, yaitu *Decision/classification* trees, *Bayesian classifiers/ Naïve Bayes classifiers*, *Neural networks*, Analisa Statistik, Algoritma Genetika, *Rough sets*, *k-nearest neighbor*, Metode *Rule Based*, *Memory based reasoning*, dan *Support vector machines* (SVM).

Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase *training*), dimana algoritma klasifikasi dibuat untuk menganalisa data *training* lalu direpresentasikan dalam bentuk *rule* klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari *rule* klasifikasi (Han, 2006). Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011) :

a. Kelas
Variabel dependen yang berupa kategorikal yang merepresentasikan ‘label’ yang terdapat pada objek. Contohnya: resiko penyakit jantung, resiko kredit, *customer loyalty*, jenis gempa.

b. *Predictor*
Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contohnya: merokok, minum alkohol, tekanan darah, tabungan, aset, gaji.

c. *Training dataset*
Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor.*

d. *Testing dataset*
Berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

Pada penelitian ini peneliti menggunakan algoritma K-Nearest Neighbor dengan pendekatan K 5, K10, K15, K20, dan untuk pengujian nya menggunakan aplikasi rapid miner. Algoritma K-NN awalnya diperkenalkan pertama kali pada awal 1950an. Metode ini tidak banyak digunakan sampai akhir 1960-an karena menggunakan sumber daya komputer yang sangat besar. Perbaikan Sumber daya computer sekarang membuat metode klasifikasi k-NN Algoritma banyak digunakan di bidang pengenalan pola a data. Klasifikasi menggunakan algoritma k-NN berdasarkan analogi, dengan membandingkan catatan tes yang diberikan oleh pelatihan catatan yang memiliki kesamaan. Catatan pelatihan digambarkan sebagai n fitur. Setiap record adalah titik dalam ruang n-dimensi. Di dalam Cara, catatan pelatihan disimpan dalam pola n-dimensi ruang. Bila ada catatan yang tidak diketahui, klasifikasi menggunakan Metode algoritma k-NN mencari pola spasial untuk a Catatan pelatihan yang sangat dekat dengan catatan. Catatan pelatihan K adalah k- "Tetangga Terdekat" dari catatan yang tidak diketahui (Arif, 2017).

Langkah-langkah untuk menghitung algoritma *k*-NN:

1. Menentukan nilai *k*.
2. Menghitung kuadrat jarak *euclid* (*query instance*) masing-masing objek terhadap *data training* yang diberikan.
3. Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak *euclid* terkecil.
4. Mengumpulkan label *class* Y (klasifikasi *Nearest Neighborhood*).
5 Dengan menggunakan kategori *Nearest Neighborhood* yang paling mayoritas maka dapat dipredeksikan nilai *query instance* yang telah dihitung (Selvia, 2014).

Untuk mengukur akurasi algoritma klasifikasi, metode yang dapat digunakan antara lain *cross validation*, *confusion matrix*, dan kurva ROC (*Receiver Operating Characteristic*). Untuk mengembangkan aplikasi (*development*) berdasarkan model yang dibuat, digunakan Rapid Miner.

a. *Cross Validation*
Cross validation adalah pengujian standar yang dilakukan untuk memprediksi *error rate*. Data *training* dibagi secara random ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan.

b. *Confusion matrix*

Metode ini menggunakan tabel matriks seperti padaTabel 1 jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007).

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negative, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity* (*recall*), *specificity*, *precision* dan *accuracy*.

Sensitivity digunakan untuk membandingkan jumlah TP terhadap jumlah *record* yang positif sedangkan *specificity* adalah perbandingan jumlah TN terhadap jumlah *record* yang negatif. Dalam perhitunganmya digunakan persamaan di bawah ini (Han, 2006) :

$$sensitivity = \frac{TP}{P}$$
$$specificity = \frac{TN}{N}$$
$$precision = \frac{TP}{TP + FP}$$
$$accuracy = sensitivity \frac{P}{(P + N)} + specificity \frac{N}{(P + N)}$$

Rumus 2. *Sensitivity*

Keterangan: TP = jumlah *true positives* TN = jumlah *true negatives* P = jumlah *record* positif N = jumlah *tupel* negatif

c. FP = jumlah *false positives*

Cross validation adalah pengujian standar yang dilakukan untuk memprediksi *error rate*. Data *training* dibagi secara random ke dalam beberapa bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan.

d. Kurva ROC

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertikal (Vecellis, 2009). *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan. AUC dihitung menggunakan rumus (Liao, 2007) :

$$\theta^{ij} = \frac{1}{n_i n_j} \sum_{x^i \in D_i} \sum_{x^j \in D_j} \psi(x^i, x^j)$$

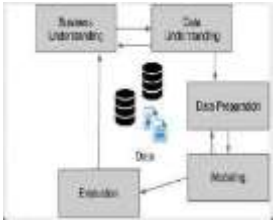
Dimana

$$(X, Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases}$$

Rumus 3. Mmenghitung kurva ROC

K = jumlah algoritma klasifikasi yang dikomparasi X = output *positif* Y = output *negative*

Untuk membantu dalam menganalisis data, peneliti menggunakan Rapid Miner. Rapid Miner merupakan perangakat lunak yang bersifat terbuka (open source). Rapid Miner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. Rapid Miner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. Rapid Miner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. Rapid Miner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. Rapid Miner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi (Ilham, 2016).



Gambar 2. Siklus CRISP-DM

Pada penelitian ini metode analisis data menggunakan metode CRISP-DM sampai proses evaluation saja.

HASIL DAN PEMBAHASAN

Business Understand

Pada siklus ini terdapat dua tahap yang dilakukan, diantaranya:

1) Tujuan Bisnis
   Pada penelitian ini memiliki tujuan untuk meningkatkan omset perusahaan dari sebelum diadakan nya penelitian.
2) Tujuan Data Mining
   Pada penelitian ini memiliki tujuan untuk memprediksi pembayaran siswa yang melakukan kursus dimasa yang akan datang setelah dilakukannya penelitian ini.

Data Understanding

Penelitian ini menggunakan 529 *record* data siswa ENTER baik yang sudah lunas maupun yang belum lunas. Semua atribut pada data *training* bernilai kategori, seperti terlihat pada Tabel 2. data *training* terdiri dari 5 atribut, dimana 4 atribut merupakan prediktor dan 1 atribut label.

Tabel 2 Daftar atribut dan nilainya

NO	ATRIBUT	NILAI ATRIBUT
1	Kategori Siswa	Mahasiswa Pelajar Umum Perusahaan
2	Sumber Biaya	Pribadi Orang tua
3	Pendapatan Pribadi	1 juta – 2 juta 2 juta – 4 juta 4 juta – 6 juta >6 juta
4	Pendapatan Orang Tua	1 juta – 2 juta 2 juta – 4 juta 4 juta – 6 juta >6 juta
5	Keterangan	Lunas Belum

Dalam pengukuran jarak antar atribut, akan diberikan bobot pada atribut. Bobot jarak ini diberikan nilai antara 0 sampai dengan 1. Nilai 0 artinya jika atribut tidak berpengaruh dan sebaliknya nilai 1 jika atribut sangat berpengaruh.

Tabel 3 pembobotan atribut

NO	ATRIBUT	BOBOT
1	Kategori Siswa	1
2	Sumber Biaya	1
3	Income Pribadi	0.5
4	Income Orang Tua	0.5

Contoh penentuan kedekatan antar nilai atribut terdapat pada table 4, misalkan untuk atribut kategori siswa terdiri dari empat nilai kategori, yaitu pelajar, mahasiswa,umum, perusahaan.

Tabel 4 kedekatan nilai atribut kategori siswa

Atribut	Nilai Atribut 1	Nilai Atribut 2	BOBOT
Kategori Siswa	Pelajar	Pelajar	0
	Pelajar	Mahasiswa	0.5
	Pelajar	Umum	1
	Pelajar	Perusahaan	1
	Mahasiswa	Mahasiswa	0
	Mahasiswa	Umum	1

	Mahasiswa	Perusahaan	1
	Umum	Umum	0
	Umum	Perusahaan	0.5
	Perusahaan	Perusahaan	0

Pembobotan nilai atribut dilakukan untuk 4 atribut prediktor. Setelah itu hitung kemiripannya. Misal sebuah data siswa baru akan diklasifikasi apakah bermasalah atau tidak dalam pembayaran biaya kurusus maka dilakukan perhitungan kedekatan antara kasus baru dibandingkan dengan data kasus lama (data *training*).

Data Preparation

Data yang berisi sampel data *training* yang merupakan kasus lama dan akan diukur kedekatannya dengan kasus yang baru.

Tabel 5 sampel data *training*

NAMA	KATSISWA	SUMBERBIAYA	INCOMEPRIBADI	INCOMEORTU	KETERANGAN
LAELA QODRIYAH P	UMUM	PRIBADI	Rp. 2 Jt - Rp. 4 Jt	-	LUNAS
MAHARANI PASHA UMAR	MAHASISWA	ORANGTUA	-	Rp. 2 Jt - Rp. 4 Jt	LUNAS
YURI PUSPITA PRATIWI	MAHASISWA	ORANGTUA	-	Rp. 2 Jt - Rp. 4 Jt	LUNAS
AMALIA ANUGRAH SUPRIYADI	UMUM	PRIBADI	Rp. 1 Jt - Rp. 2 Jt	-	BELUM
MUHAMMAD ALDIN SUPRIYADI	UMUM	PRIBADI	Rp. 1 Jt - Rp. 2 Jt	-	BELUM

Misalkan ada kasus baru pada data *testing* dengan nilai atribut seperti pada tabel 6. Kasus baru tersebut akan dihitung kedekatannya dengan kasus lama yang terdapat pada data training table 5.

Tabel 6 sampel data *testing*

NAMA	KATSISWA	SUMBERBIAYA	INCOMEPRIBADI	INCOMEORTU	KETERANGAN
NURADILLA ZAHRA	MAHASISWA	ORANGTUA	-	Rp. 2 Jt - Rp. 4 Jt	LUNAS

Perhitungan kedekatan kasus baru pada data *testing* (Tabel 6) dengan 5 kasus lama pada data *training* (Tabel 5), yaitu: Kedekatan kasus baru dengan kasus nomor 1 dimana kedekatan bobot atribut katsiswa (umum dengan mahasiswa) = 1, bobot atribut status Katsiswa = 1, kedekatan bobot sumberbiaya (pribadi dengan orang tua) = 1, bobot atribut sumber biaya = 1, kedekatan bobot income pribadi (2 jt – 4 jt dengan 2 jt – 4 jt) = 0, bobot atribut income pribadi = 0.5, kedekatan bobot income orang tua (2 jt – 4 jt dengan 2 jt – 4 jt) = 0, bobot atribut income orang tua = 0.5

Similarity = [(A*B) + (C*D) + (E*F) + (G*H)/ (B+D+F+H)]

= [(1*1) + (1*1)+(0*0.5)+ (0*0.5)/(1+1+0.5+0.5)]

=(1+1+0+0)/3

= 2/3 = 0.67

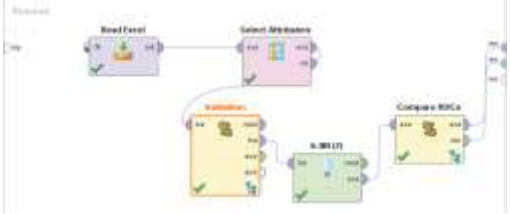
Lakukan perhitungan diatas sampai dengan k5 sehingga di hasilkan seperti di tabel 6.

Kedekatan 1	Kedekatan 2	Kedekatan 3	Kedekatan 4	Kedekatan 5
0.67	0	0	1.33	1.33

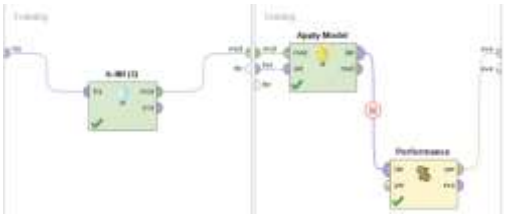
Setelah dihitung nilai kedekatannya yang terendah adalah kasus nomor 2 dan 3. Dengan demikian kasus yang terdekat dengan kasus baru adalah kasus nomor 1. Jadi kemungkinan siswa baru tersebut akan membayar lunas biaya kursus.

Modeling

Model algoritma yang digunakan pada penelitian prediksi pembayaran siswa kursus menggunakan algoritma K-NN dimana, hasil yang diharapkan dalam bentuk gambar dan rule menggunakan perangkat lunak RapidMiner 8.0. Desain algoritma K-NN pada RapidMiner dijelaskan pada gambar 3 dan gambar 4



Gambar 3. Proses Model KNN



Gambar 3. Deklarasi dari Model KNN

Evaluation

Evaluasi dilakukan melalui pengujian algoritma, dengan beberapa tahap, diantaranya:

1. *Cross Validation*

Dalam penelitian ini digunakan *10 fold-cross validation* dimana 529 record pada data *training* dibagi secara random ke dalam 10 bagian dengan perbandingan yang sama kemudian *error rate* dihitung bagian demi bagian, selanjutnya hitung rata-rata seluruh *error rate* untuk mendapatkan *error rate* secara keseluruhan.

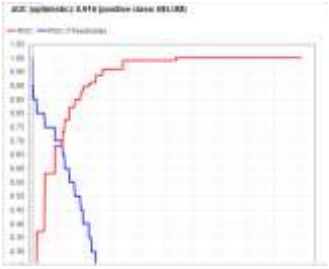
2. *Confusion Matrix*

Tabel 8 adalah table *confusion matrix* yang dihasilkan dengan menggunakan algoritma kNN. Perhitungan kedekatan kasus lama pada data *training* dengan kasus baru pada data *testing*, berikut data confusion matrix berdasarkan K5, k10, k15, dan k20

K5	K10	K15	K20
84.69%	83.93%	83.36%	83.36%

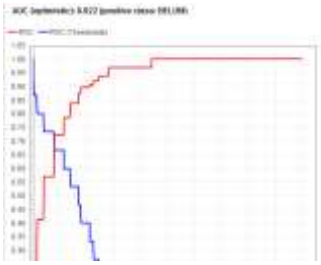
Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Kurva ROC pada K20 mengekspresikan *confusion matrix* dari Tabel 8. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Terlihat pada table, nilai AUC sebesar 0.916.



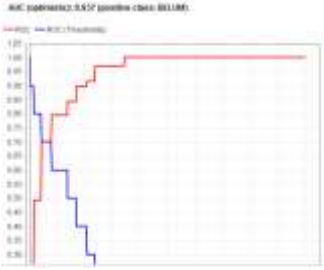
Gambar 2. Kurva ROC Pada Pendekatan K 20

Hasil perhitungan divisualisasikan dengan kurva ROC. Kurva ROC pada K15 mengekspresikan *confusion matrix* dari Tabel 8. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Terlihat pada table, nilai AUC sebesar 0.922.



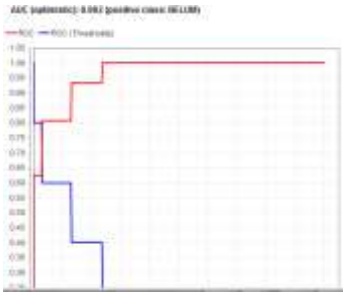
Gambar 3. Kurva ROC Pada Pendekatan K 15

Hasil perhitungan divisualisasikan dengan kurva ROC. Kurva ROC pada K10 mengekspresikan *confusion matrix* dari Tabel 8. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Terlihat pada table, nilai AUC sebesar 0.937.



Gambar 4. Kurva ROC Pada Pendekatan K 10

Hasil perhitungan divisualisasikan dengan kurva ROC. Kurva ROC pada K5 mengekspresikan *confusion matrix* dari Tabel 8. Garis horizontal adalah *false positives* dan garis vertikal *true positives*. Terlihat pada table, nilai AUC sebesar 0.962.



Gambar 5. Kurva ROC Pada Pendekatan K 5

Dalam klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = klasifikasi sangat baik
b. 0.80-0.90 = klasifikasi baik
c. 0.70-0.80 = klasifikasi cukup
d. 0.60-0.70 = klasifikasi buruk
e. 0.50-0.60 = klasifikasi salah

Berdasarkan pengelompokkan di atas maka dapat disimpulkan bahwa metode kNN pada prediksi pembayaran siswa lembaga kursus ENTER termasuk klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

KESIMPULAN

Dalam penelitian ini dilakukan penerapan algoritma kNN pada data siswa yang melakukan pembayaran biaya kursus di ENTER (English Center). Agar didapat data yang berkualitas, dilakukan *preprosesing* sebelum diterapkan ke dalam algoritma. Kedekatan antara kasus baru dengan kasus lama dilakukan untuk menentukan termasuk kelas mana kasus baru tersebut. Untuk mengukur kinerja algoritma tersebut digunakan metode *Cross Validation*, *Confusion Matrix* dan Kurva ROC k5, k10, k15, k20 berturut-turut 84.69%, 83.93%, 83.36%, 83.36% dan menghasilkan akurasi dan nilai AUC 0.962, 0.937, 0.922, 0.916. Karena nilai AUC berada dalam rentang 0,9 sampai 1,0 maka metode tersebut masuk dalam kategori sangat baik.

DAFTAR PUSTAKA

Antaranews. 2016. Indonesia Peringkat 32 Penguasaan bahasa Inggris, https://www.antaranews.com/berita/600584/indonesia-peringkat-32-penguasaan-bahasa-inggris. (Diakses pada:11 Januari 2018)

Bramer, Max. 2007. *Principles of Data Mining*. London : Springer

Ginting L Selvia. 2014. Teknik Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighborhood. Bandung

Gorunescu, Florin. 2011. *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg : Springer

Ilham R Mochamad. 2016. Implementation Of Data Mining Using C4.5 Algorithm For Prediction Of Customer Satisfaction At Kosti Taxi. Semarang

Kementrian Pendidikan. 2003. Putusan Kementrian Pendidikan Republik Indonesia

Larose, D. T. 2005. *Discovering Knowledge in Data*. New Jersey : John Willey & Sons, Inc.

Liao. 2007. *Recent Advances in Data Mining of Enterprise Data* : *Algorithms and Application*. Singapore : *World Scientific Publishing*

Maimon, Oded&Rokach, Lior. 2005. *Data Mining and Knowledge Discovey Handbook*. New York : Springer

Sumathi, & S., Sivanandam, S.N. 2006. *Introduction to Data Mining and its Applications*. Berlin Heidelberg New York: Springer

Syarif R Arif. 2017. Intrusion Detection System Using Hybrid Binary Pso And K-Nearest Neighborhood Algorithm. Jakarta

Witten, I. H., Frank, E., & Hall, M. A. 2011. *Data Mining: Practical Machine Learning and Tools*. Burlington : Morgan Kaufmann Publisher

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz,C. Shearer, & R. Wirth, *CRISP-DM 1.0 Step-by-step data mining guide*, 2000