Pelita Teknologi: Jurnal Ilmiah Informatika, Arsitektur dan Lingkungan
Journal homepage: jurnal.pelitabangsa.ac.id

PENENTUAN KELULUSAN SISWA YAYASAN CERDAS BAKTI PERTIWI DENGAN MENGGUNAKAN ALGHORITMA *NAÏVE BAYES* DAN *CROSS VALIDATION*

Elkin Rilvani, Ahmad Budi Trisnawan, Priasnyomo Prima Santoso

Program Studi Teknik Informatika Sekolah Tinggi Teknologi Pelita Bangsa
Korespondensi email: *elkin.rilvani@pelitabangsa.ac.id*

Abstrak

Yayasan Cerdas Bakti Pertiwi sebagai unit pelaksana pendidikan non formal dalam mencapai tujuan pendidikan dan pelatihan menyiapkan peserta didik untuk kedinasan bintara yang bisa menjadi penerus bangsa untuk dapat menjawab tantangan zaman. Dalam kegiatan operasionalnya siswa dituntut agar lulus namun untuk mencapai kelulusan banyak faktor yang menjadi tantangan hambatan siswa. Pada penelitian ini akan memecahkan permasalahan faktor hambatan kelulusan dengan data mining untuk menentukan kelulusan siswa. Teknik data mining adalah klasifikasi dengan metode Naïve Bayes dan Cross Validation, maka didapatkan hasil penentuan kelulusan siswa dengan persentase keakuratan sebesar 99,4 %. Dalam penelitian menggunakan data sebanyak 500 siswa yang terdiri dari 443 siswa laki-laki dan 57 siswa perempuan.

Informasi Artikel

Diterima : 02 September 2019
Direvisi : 05 September 2019
Dipublikasikan: 09 September 2019

Kata kunci

Pendidikan Non Formal, Data Mining, Naïve Bayes dan Cross Validation

1. PENDAHULUAN

1.1 Latar Belakang

Yayasan Cerdas Bakti Pertiwi sebagai unit pelaksana pendidikan non formal bergerak dalam bimbingan belajar untuk kedinasan bintara dan diharapkan mampu melaksanakan fungsi serta tugasnya dengan baik. Seiring perkembangan zaman yang semakin dinamis maka institusi pendidikan non formal dituntut untuk dapat menyesuaikan diri dalam rangka menyiapkan peserta didik yang bisa menjadi penerus bangsa untuk dapat menjawab tantangan zaman. Dalam kegiatan operasionalnya didapatkan data yang berlimpah mengenai para siswa dalam proses pembelajaran dan pelatihan untuk mencapai kelulusannya yang tersimpan dalam basis data (database).

Klasifikasi merupakan salah satu teknik data mining yang digunakan untuk membangun suatu model dari sampel data yang belum terklasifikasi untuk digunakan mengklasifikasi sampel data baru ke dalam kelas-kelas yang sejenis. Pemilihan alghoritma klasifikasi yang tepat untuk prediksi service adalah hal yang sangat mempengaruhi kepercayaan pelanggan kepada perusahaan oleh karena itu penulis memilih Naive Bayes untuk peneltian ini "Naive Bayes adalah salah satu algoritma yang paling banyak digunakan dalam masalah klasifikasi karena sederhana dan keefektifannya. Ini cocok untuk banyak skenario pembelajaran, seperti klasifikasi gambar, deteksi penipuan, penambangan web, dan klasifikasi teks (Arar and Ayan, 2017). Namun Naive Bayes juga memiliki masalah dalam akurasi perhitungan hasilnya "Pengklasifikasian Naive Bayes adalah salah satu metode klasifikasi yang sederhana namun kuat, tetapi memiliki permalasahan pada kelas-kelas crips yang ditugaskan untuk data training (Karabatak, 2015). Pada permasalah Naive Bayes ini akan dibantu Cross Validation untuk meningkatkan akurasi perhitungannya "Cross Validation adalah pendekatan yang telah dicoba dan diuji untuk memilih model yang optimal (Liu and Liao, 2017).

1.2 Rumusan Masalah

Berdasarkan identifikasi masalah diatas adalah agar dapat menentukan tingkat kelulusan dengan data siswa, maka pertanyaan penelitian yang timbul dalam penelitian adalah, "Bagaimana menentukan tingkat kelulusan siswa menggunakan algoritma naïve bayes dan cross validation dengan akurasi yang optimal ?".

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini menggunakan data mining untuk menentukan tingkat kelulusan siswa pada Yayasan Cerdas Bakti Pertiwi agar dapat membantu :

a. Mengetahui faktor-faktor yang sangat berpengaruh pada tingkat kelulusan.
b. Mengoptimalkan kegiatan pembelajar dan pelatihan siswa berpengaruh pada tingkat kelulusan.

2. METODE PENELITIAN

Metode yang digunakan dalam penyusunan penelitian dilakukan dengan beberapa tahap yaitu:

a. Explorasi dan studi literature
   Dengan diadakannya studi mengenai data mining, dengan menerapkan metode naïve bayes dan cross validation pada klasifikasi kelulusan, melalui literature seperti jurnal, buku, artikel di halaman web.
b. Observasi dan Wawancara
   Pengumpulan data dilakukan dengan pengamatan langsung dan wawancara dengan beberapa narasumber
c. Pengimplementasian dalam aplikasi
   Setelah proses data mining dilakukan sesuai tahapnya maka perhitungan dilakukan dalam aplikasi untuk memvalidasi tingkat akurasi klasifikasi.
d. Kesimpulan

Menganalisa secara keseluruhan hasil penelitian mengambil kesimpulan dari data-data yang didapat.

2.1 Pengertian Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database. Data mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan, dan mechine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Sharda, Delen and Turban, 2014).

2.2 Konsep Klasifikasi

Klasifikasi adalah salah satu pembelajaran yang paling umum di data mining. Klasifikasi didefinisikan sebagai bentuk analisis data untuk mengekstrak model yang akan digunakan untuk memprediksi label kelas (Sartika *et al.*, 2017)

Klasifikasi adalah salah satu tugas dari data mining yang bertujuan untuk memprediksi label kategori benda yang tidak diketahui sebelumnya, dalam membedakan antara objek yang satu dengan yang lainnya berdasarkan atribut atau fitur (Mutrofin, 2014).

2.3 Naïve Bayes

Algoritma naïve bayes adalah salah satu algoritma yang tergolong kedalam statistical classifier algoritma tersebut pertama kali diperkenalkan oleh Thomas Bayes dimana algoritma ini merupakan adaptasi dari theorem bayes (Han, Kamber and Pei, 2012). Algoritma ini mengandalkan sebuah peluang kemungkinan suatu objek.

Rumus dari *theorema bayes:*

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Keterangan :

X = Data dengan class yang belum diketahui

H = Hipotesis data X merupakan suatu class spesifik

P(H|X)= Probabilitas hipotesis H berdasarkan kondisi X (*posteriori prob*)

P(H) = Probabilitas hipotesis H *(prior prob*)

P(X|H) =Probabilitas X berdasarkan kondisi tersebut

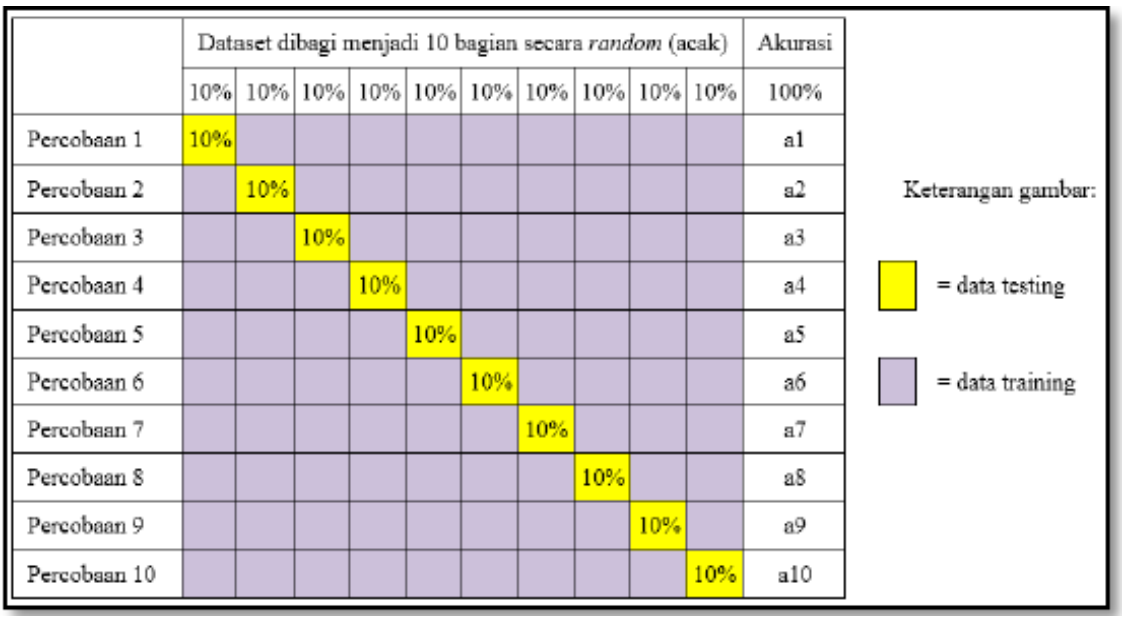
P(X) =Probabilitas dari X

Naive Bayes adalah metode yang digunakan dalam statistika untuk menghitung peluang dari suatu hipotesis, Naive Bayes menghitung peluang suatu kelas berdasarkan pada atribut yang dimiliki dan menentukan kelas yang memiliki probabilitas paling tinggi. Naive Bayes mengprediksikan kelas berdasarkan pada probabilitas sederhana dengan mangasumsikan bahwa setiap atribut dalam data tersebut bersifat saling terpisah. Metode Naive Bayes merupakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana, metode Naive Bayes mengprediksikan data berdasarkan probabilitas P atribut x dari setiap kelas y data (Socrates, Akbar and Akbar Sonhaji, 2016).

2.4 Cross Validation

Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua *subset* yaitu data proses pembelajaran dan data *validasi* / evaluasi *Cross validation* adalah membagi dataset menjadi dua bagian dengan satu bagian dijadikan data *training* dan bagian yang lain dijadikan data *testing*. Beberapa penelitian membagi data menjadi 10 bagian, 90% dijadikan *training* dan 10 lainnya digunakan sebagai *testing.* Proses ini dilakukan berulang sampai dengan 10 kali hingga semua *record* data mendapatkan bagian menjadi data *testing*. Proses ini dikenal juga dengan istilah *10 folds cross validation. 10 folds cross validation* banyak digunakan peneliti karena terbukti menghasilkan

performa algoritma yang lebih stabil. Gambar 1 merupakan representasi dari *10 folds cross validation* (Indrayanti, Sugianti and Karomi, 2017).



Gambar 2.1 Skema 10 Fold CV

3. HASIL DAN PEMBAHASAN

3.1 Teknik Analisis

Dalam penelitian ini menggunakan model CRISP-DM (Cross Standart Industries for Data Mining), yang memiliki siklus hidup terdiri dari beberapa tahap. Keseluruhan fase berurutan yang ada dan bersifat adaptif. Fase berikutnya dalam urutan bergantung kepada keluaran dari fase sebelumnya. Adapun tahapan-tahapannya sebagai berikut :

3.1.1 Business Understanding

Pada tahapan pertama penulis mencoba untuk memahami permasalahan yang ada dalam penentuan kelulusan pada yayasan cerdas bakti pertiwi. Sehingga dapat menentukan tujuan dan pola yang akan didapatkan dengan data mining.

Faktor dalam penentuan kelulusan siswa bisanya terdapat pada salah satu atribut yang mempunyai faktor terbesar dalam menentukan kelulusan, dimana atribut berperan terhadap kelulusan dan yayasan, sehingga hasil kelulusan pada beberapa daerah dikembalikan lagi terhadap yayasan yang bersangkutan

3.1.2 Data Understanding Phase (Fase Pemahaman Data)

Pada tahap ini penulis melakukan pemahaman terhadap data yang dibutuhkan, untuk kemudian mengambil data yang relevan dan memiliki keterkaitan dengan tujuan penelitian. Adapun data yang digunakan yaitu data hasil penilaian siswa yang memiliki beberapa atribut diantaranya: “nilai evaluasi, nilai smapta, hasil medical check up, jenis kelamin dan keterangan”.

3.1.3 Data Preparation Phase (Fase Pengolahan Data)

Pada proses pembersihan data adalah proses untuk membersihkan data yang dihasilkan pada tahapan mengvaluasi data. Pada tahap pembersihan data ini melakukan pembersihan data sebagai berikut :

a. Tahap Pertama, penentuan data yang akan diolah pada penelitian ini Berikut pada tabel 3.2 merupakan tabel atribut data penelitian dari dataset.

Tabel 3.1 Atribut Dataset

No	Atribut	*Type*
1	Nama Siswa dan Siswi	*Char*
2	Nilai Evaluasi	*Varchar*
3	Nilai Smapta	*Varchar*
4	Medical Check Up	*Char*
5	Jenis Kelamin	*Char*
6	Keterangan	*Char*

b. Tahap Kedua, melakukan konversi data. Data dengan atribut yang telah dipilih kemudian dikonversikan untuk memudahkan proses data mining pada sebagian atribut, karena data akan diproses dengan *tools* bantu data *mining*.

3.1.4 *Modeling Phase* (Fase Pemodelan)

Pada tahap ini penulis menentukan teknik *data mining* yang digunakan untuk mengolah data yang sudah disiapkan sebelumnya. Teknik yang dilakukan yaitu dengan *klasifikasi* menggunakan algoritma *Naïve Bayes dan Cross Validation*. Data yang sudah melalui proses pengolahan kemudian akan dilakukan perhitungan dengan menggunakan *tools Rapidminer*. Dua langkah yang dilakukan pada tahap ini ialah :

a. Perhitungan *Naïve Bayes* dan *Cross Validation* secara manual Data yang akan digunakan dalam perhitungan *Naïve Bayes* secara manual yaitu 500 sampel data. Dimana data yang diambil secara acak oleh peneliti.

Tabel 3.2 Menghitung penentuan kelulusan atau prediksi

No	P(Ci)	PCI/X	P(X|C1)	P(X|C2)	P(X|C3)	P(X|C4)	P(X|Ci)	P(X|Ci)*P(CI)	keterangan
1	485	0.97	0.076	1	0.301	0.887	0.0204	0.0198	LULUS
	15	0.03	0	1	0.2	0.867	0	0	
2	485	0.97	0.076	1	0.699	0.887	0.0473	0.0459	LULUS
	15	0.03	0	1	0.8	0.867	0	0	
3	485	0.97	0.058	1	0.301	0.887	0.0154	0.0149	LULUS
	15	0.03	0	1	0.2	0.867	0	0	
4	485	0.97	0.058	1	0.699	0.887	0.0358	0.0347	LULUS
	15	0.03	0	1	0.8	0.867	0	0	
5	485	0.97	0.037	1	0.699	0.887	0.023	0.0223	LULUS
	15	0.03	0	1	0.8	0.867	0	0	
..	..	..	..	..	..	..	..		..
..	..	..	..	..	..	..	..		..
496	485	0.97	0.076	1	0.301	0.113	0.0026	0.0025	LULUS
	15	0.03	0	1	0.2	0.133	0	0	
497	485	0.97	0.045	1	0.301	0.113	0.0015	0.0015	LULUS
	15	0.03	0	1	0.2	0.133	0	0	
498	485	0.97	0.045	1	0.699	0.113	0.0036	0.0035	LULUS
	15	0.03	0	1	0.8	0.133	0	0	
499	485	0.97	0.039	1	0.699	0.113	0.0031	0.003	LULUS
	15	0.03	0	1	0.8	0.133	0	0	
500	485	0.97	0.058	1	0.699	0.113	0.0046	0.0044	LULUS
	15	0.03	0	1	0.8	0.133	0	0	

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

b. Perhitungan *Cross Validation*

Pada penelitian ini perhitungan cross validation dilakukan sebanyak 10 kali dan didapatkan hasil sebagai berikut :

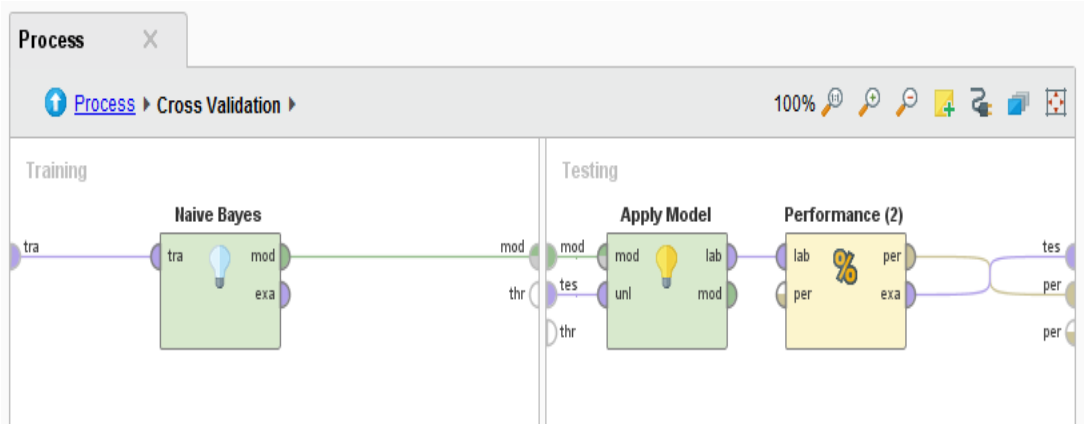
Tabel 3.3 Cross Validation K-Fold 10

No	Nilai Evaluasi	Nilai Smapta	HASIL MEDICAL CHECK UP	Jenis Kelamin	Keterangan
451	73	MS	TMS	LAKI -LAKI	LULUS
452	84	MS	TMS	LAKI -LAKI	LULUS
453	69	MS	TMS	LAKI -LAKI	LULUS
..	..	..	..	..	..
..	..	..	..	..	..
498	79	MS	TMS	PEREMPUAN	LULUS
499	78	MS	TMS	PEREMPUAN	LULUS
500	75	MS	TMS	PEREMPUAN	LULUS

Perhitungan Cross Validation		
Confusion Matrix Fold 10		
Prediksi	LULUS	TIDAK LULUS
LULUS	48	0
TIDAK LULUS	0	2

3.2 Hasil Perhitungan Akurasi

Model pengujian dimana pada dataset source nantinya akan disesuaikan dengan dataset nilai siswa. Kemudian untuk memvalidasi model dari algoritma naïve bayes digunakan metode cross validation. Dimana didalamnya terdapat performance dengan menggunakan confussion matrix sebagai model evaluasi dari kinerja algoritma naïve bayes. Untuk lebih jelasnya dapat dilihat pada gambar 4.4 berikut :



Gambar 3.1 Naïve Sub Proses Cross Validation

Dalam sub proses *cross validation* terdapat dua bagian dimana ada *training* dan juga *testing* seperti pada di gambar 3.1 dimana pada bagian *training* terdapat algoritma *naïve bayes* hal tersebut dimaksudkan agar dataset dibuat modelnya menggunakan algoritma *naïve bayes* dimana pada *cross validation* dengan nilai k = 10 *folds* sehingga membagi data menjadi 10:90 yakni 10% dijadikan data testing dan 90% merupakan data training. Kemudian pada data testing terdapat dua fitur yakni *Apply model* yang digunakan untuk menerapkan model data yang sudah dilatih sebelumnya dengan data *test.*

Terakhir dibagian *testing* terdapat fitur *performance* dimana fitur tersebut digunakan untuk mengevaluasi hasil kinerja algoritma *naïve bayes* dengan parameter pengukuran *confussion matrix* (*accuracy,recall,precision*).

Pengujian dengan metode Naïve Bayes menggunakan dataset Nilai Siswa. Hasil yang didapatkan pengujian ini mendapatkan hasil akurasi sebesar 99,40% dengan nilai presisi serta recall masing-masing kelas dapat dilihat pada gambar 4.7 berikut ini :

accuracy: 99.40% +/- 0.92% (mikro: 99.40%)

	true LULUS	true TIDAK LULUS	class precision
pred. LULUS	482	0	100.00%
pred. TIDAK LULUS	3	15	83.33%
class recall	99.38%	100.00%	

Gambar 3.2 Confusion Matrix Nilai Siswa

Hasil analisa antara data training dan data test pada *rapidminer* dapat untuk menghitung akurasi sebagai berikut:

Diketehui :
Jumlah data yang diuji : 500
Jumlah data yang diprediksi benar : 497
Jumlah data yang diprediksi salah : 3

$$Akurasi = \frac{\Sigma \text{data uji benar klasifikasi}}{\Sigma \text{total data uji}} \times 100\%$$

Hitungan Hasil Akurasi

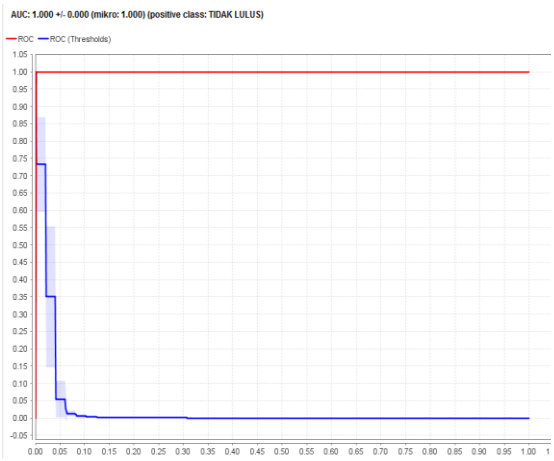
$$Akurasi = \frac{497}{500} x100\%$$

$$\text{Hasil} = 99{,}4\%$$

3.3 Pembahasan

Penelitian yang dilakukan menggunakan algoritma *naïve bayes* dan *cross validation* menentukan *Area Under Curve (AUC)* dan pola baru dalam kelulusan/faktor-faktor yang berpengaruhi kelulusan

3.3.1 Area Under Curve (AUC)



Gambar 3.3 Kurva ROC

Kurva ROC digunakan untuk mengekspresikan data, garis horizontal mewakili nilai *false positive* dan garis vertical mewakili nilai *true positive.* Dari gambar 3.3 dapat diketahui bahwa nilai *Area Under Curve (AUC)* model algoritma *naïve bayes* dan c*ross validation* adalah 1.000,hal ini menunjukan bahwa algoritma digunakan mencapai klasifikasi sempurna.

3.2 Pola Penentu Kelulusan

Berdasarkan proses data mining penentuan kelulusan siswa dalam penelitian ini dapat diperoleh informasi bahwa pada hasil penilaian kelulusan berdasarkan label "Lulus dan Tidak Lulus" berdasarkan atribut nilai sebagai berikut :

a. LULUS :

Nilai Evaluasi : 61 sampai 91, Nilai Smapta : MS(Memenuhi Syarat), MCU : MS(Memenuhi Syarat dan Jenis Kelamin : Laki-laki

b. TIDAK LULUS :

Nilai Evaluasi : 47 sampai 60, Nilai Smapta : MS(Memenuhi Syarat), MCU : TMS(Tidak Memenuhi Syarat dan Jenis Kelamin : Perempuan

1. Pola pertama terdiri dari atribut nilai dimana nilai evaluasi 61; nilai smapta MS; MCU TMS; dan Jenis Kelamin laki-laki; mendapatkan keterangan Lulus.
2. Pola kedua terdiri dari atribut nilai dimana nilai evaluasi 91; nilai smapta MS; MCU TMS; dan Jenis Kelamin laki-laki; mendapatkan keterangan Tidak Lulus.
3. Pola ketiga terdiri dari atribut nilai dimana nilai evaluasi 60; nilai smapta MS; MCU TMS; dan Jenis Kelamin Perempuan; mendapatkan keterangan Lulus.
4. Pola keempat terdiri dari atribut nilai dimana nilai evaluasi 80; nilai smapta MS; MCU TMS; dan Jenis Kelamin Perempuan; mendapatkan keterangan Tidak Lulus.
5. Pola kelima terdiri dari atribut nilai dimana nilai evaluasi 69; nilai smapta MS; MCU TMS; dan Jenis Kelamin Perempuan; mendapatkan keterangan Lulus
6. Pola Keenam terdiri dari atribut nilai dimana nilai evaluasi 58; nilai smapta MS; MCU TMS; dan Jenis Kelamin laki-laki; mendapatkan keterangan Lulus.
7. Pola ketujuh terdiri dari atribut nilai dimana nilai evaluasi 60; nilai smapta MS; MCU TMS; dan Jenis Kelamin

laki-laki; mendapatkan keterangan Tidak Lulus.

8. Pola kedelapan terdiri dari atribut nilai dimana nilai evaluasi 47; nilai smapta MS; MCU TMS; dan Jenis Kelamin laki-laki; mendapatkan keterangan Lulus.
9. Pola kesembilan terdiri dari atribut nilai dimana nilai evaluasi 73; nilai smapta MS; MCU TMS; dan Jenis Kelamin laki-laki; mendapatkan keterangan Tidak Lulus.
10. Pola kesepuluh terdiri dari atribut nilai dimana nilai evaluasi 72; nilai smapta MS; MCU TMS; dan Jenis Kelamin Perempuan; mendapatkan keterangan Lulus.

4. KESIMPULAN

Berdasarkan data siswa yang diperoleh, proses data mining dengan metode klasifikasi menggunakan alghoritma Naive Bayes dan Cross Validation didapatkan informasi dari hasil penentuan kelulusan siswa pada data pelatihan pada Yayasan Cerdas Bakti Pertiwi yang dijadikan data training, Sehingga dengan demikian metode Naive Bayes dan Cross Validation ini berhasil mengklasifikasi kelulusan siswa dengan persentase keakuratan sebesar 99,4 % dan ini merupakan excellent classification (Andriani, A. 2013)

DAFTAR PUSTAKA

Andriani, A. (2013). Sistem Pendukung Keputusan Berbasis Decision Tree Dalam Pemberian Beasiswa Studi Kasus : AMIK " BSI Yogyakarta ," *2013*(Sentika).

Arar, Ö. F. and Ayan, K. (2017) 'A feature dependent Naive Bayes approach and its application to the software defect prediction problem', *Applied Soft Computing Journal*. Elsevier B.V., 59, pp. 197–209. doi: 10.1016/j.asoc.2017.05.043.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*, *San Francisco, CA, itd: Morgan Kaufmann*. doi: 10.1016/B978-0-12-381479-1.00001-0.

Indrayanti, Sugianti, D. and Karomi, M. A. Al (2017) 'Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus', in, pp. 551–554.

Karabatak, M. (2015) 'A new classifier for breast cancer detection based on Naïve Bayesian', *Measurement: Journal of the International Measurement Confederation*. Elsevier Ltd, 72, pp. 32–36. doi: 10.1016/j.measurement.2015.04.028.

Liu, Y. and Liao, S. (2017) 'Granularity selection for cross-validation of SVM', *Information Sciences*. Elsevier Inc., 378, pp. 475–483. doi: 10.1016/j.ins.2016.06.051.

Mutrofin, S. (2014) 'Optimasi teknik klasifikasi modified k nearest neighbor menggunakan algoritma genetika', *JURNAL GAMMA*, (September), pp. 130–134.

Sartika, D. *et al.* (2017) 'Perbandingan Algoritma Klasifikasi Naive Bayes , Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian', 1(2), pp. 151–161.

Sharda, Delen and Turban, E. (2014) *Business Intelligence and Analytics*.

Socrates, Akbar and Akbar Sonhaji (2016) 'Optimasi Naive Bayes Dengan Pemilihan Fitur Dan Pembobotan Gain Ratio', *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 7(1), p. 22. doi: 10.24843/LKJITI.2016.v07.i01.p03.