PREDIKSI TINGKAT KEBERHASILAN STUDI KINERJA SANTRI MENGGUNAKAN ALGORITMA C 5.0

Achmad Agus Athok Miftachuddin.*, Kusrini*, Emha Taufiq Luthfi*

Magister Teknik Informatika, Universitas Amikom Yogyakarta

Correspondence Author: agusathok7@gmail.com

Info Artikel :

Sejarah Artikel :

Menerima :
19 Feb 2020
Revisi :
28 Mei 2020
Diterima :
14 Juni 2020
Online :
16 Agust 2020

Keyword :
Students Performance, Data mining, Prediction, C 5.0 Algorithm

ABSTRACT

The success of pesantren education institutions can be measured by the success of their students. By predicting the possible outcomes of the learning process based on prediction results can help an Islamic boarding school, by adjusting the factors that contribute and influence the success rate of students' performance studies. And by utilizing data mining techniques that can be used to increase the level of success and reduce the failure of students. this can greatly help pesantren educational institutions to improve their graduates 'skills, because data mining is the best solution to find hidden patterns and can predict the success of students' performance studies. This research presents a model based on decision tree classification algorithm C 5.0 used in this model with alumni tracer study filled by santri alumni. In this study also used the k-folds cross validation test scenario with k values of 2,3,6,10 and 15 with a total of 300 alumni data and 84 data used for validation tests without cross validation. Determination of the criteria for the classification results using a confusion matrix form the measurement of the classification results obtained, namely the highest value in this study is 95% resulting from 15 folds the scenario 1. And form the results of testing the validation data without cross validation, the corresponding results are 73.81%, when compared to the k-folds, there was an increase of 21.19% and it can be ignored that the C 5.0 algorithm is able to classify well. So that pesantren educational institutional can provide a foundation in the arrangement for their students in deciding the right school choice.

Kata Kunci :
Kinerja Santri, Data Mining , Prediksi, Algoritma C 5.0

INTISARI

Keberhasilan lembaga pendidikan pesantren dapat diukur dari keberhasilan santrinya. Dengan memprediksi kemungkinan hasil dari proses pembelajaran berdasarkan hasil prediksi dapat membantu suatu lembaga pendidikan pesantren, dengan menyesuaikan faktor-faktor yang berkontribusi dan mempengaruhi tingkat keberhasilan studi kinerja santri. Dan dengan memanfaatkan teknik data mining yang dapat digunakan untuk meningkatkan tingkat keberhasilan dan mengurangi kegagalan santri. hal ini dapat sangat membantu lembaga pendidikan pesantren untuk meningkatkan kecakapan lulusannya, karena data mining merupakan solusi terbaik untuk menemukan pola tersembunyi dan dapat memprediksi tingkat keberhasilan studi kinerja santri. Penelitian ini menyajikan model berdasarkan pohon keputusan klasifikai algoritma C 5.0 yang digunakan dalam model ini dengan tracer study online yang diisi oleh alumni santri. Pada penelitian ini juga menggunakan skenario uji k-folds cross validation dengan nilai k yaitu 2, 3, 6, 10 dan 15 dengan total 300 data alumni dan 84 data digunakan untuk uji validasi tanpa cross validation. Penentuan kriteria pada hasil klasifikasi menggunakan confusion matrix dari pengukuran hasil klasifikasi di peroleh hasil yaitu nilai akurasi tertinggi pada penelitian ini adalah 95% yang

ISSN Print : 1979-7141
ISSN Online : 2541-1942

	dihasilkan dari 15 folds skenario 1. Dan dari hasil pengujian data validasi tanpa cross validation diperoleh hasil akurasi sebasar 73,81%, jika dibandingkan dengan k-folds maka terjadi peningkatan sebesar 21,19% dan dapat disimpulkan bahwa algoritma C 5.0 mampu melakukan pengklasifikasian dengan baik. sehingga lembaga pendidikan pesantren dapat menjadikan landasan dalam pengaturan bagi santrinya dalam memutuskan pilihan sekolah yang tepat.

1. PENDAHULUAN

Pesantren merupakan sebuah pendidikan tradisional yang para santrinya tinggal bersama dan belajar dibawah bimbingan guru yang lebih dikenal dengan sebutan *kiai* dan mempunyai *asrama* untuk tempat menginap santri. Santri tersebut berada dalam asrama yang juga menyediakan masjid untuk beribadah, ruang untuk belajar, dan kegiatan keagamaan lainnya. Asrama ini biasanya dikelilingi oleh tembok untuk dapat mengawasi keluar masuknya para santri sesuai dengan peraturan yang berlaku (Dhofier, 1994).

Secara etimologi, istilah pondok pesantren berasal dari kata Bahasa arab *funduk*, dan santri yang diberi imbuhan *per* dan *an*. Kata *funduk* berarti ruang tidur atau wisma sederhana. Sedangkan kata *pesantren* berarti tempat para *santri*. Kata *"santri"* juga diartikan sebagai penggabungan antara suku kata *sant* yang berarti manusia baik dan *tra* yang berarti suka menolong sehingga kata *pesantren* dapat diartikan sebagai tempat mendidik manusia (Idoochi Anwar, 2004: 102).

Di *pesantren* selain untuk mempelajari ilmu agama Islam lebih mendalam para *santri* juga diwajibkan mengikuti pendidikan formal yaitu sekolah yang merupakan lembaga pendidikan yang memiliki tanggung jawab untuk memberi pengetahuan, keterampilan dan mengembangkannya dalam bentuk kegiatan sekolah. Sekolah dan nyantri adalah solusi untuk memperoleh keseimbangan ilmu pengetahuan.

Permasalahan yang dihadapi para *santri* baru yang berasal dari jauh adalah terdapat banyaknya pilihan sekolah yang dapat membingungkan para *santri* dalam menentukan sekolah yang sesuai sehingga *santri* mengalami kesulitan untuk mendapatkan data dan informasi secara lengkap. Oleh karena itu *santri* baru harus benar-benar mempertimbangkan dalam menentukan sekolah yang sesuai sebelum mengambil keputusan.

Sekolah mempunyai peranan penting dalam meningkatkan kecakapan lulusan dan menyiapkan lulusan untuk memasuki lapangan kerja serta mengembangkan sikap profesional, sekolah juga menyiapkan lulusan agar mampu meniti karir, mampu berkompetisi dan mampu mengembangkan diri, sekolah berusaha menyiapkan lulusan agar menjadi warga negara yang produktif, adaptif dan kreatif. Maka lembaga pendidikan khususnya sekolah memiliki tanggung jawab yang sangat relevan terhadap pembentukan jiwa *entrepreneurship* bagi lulusannya (Wahyuni and Hidayati, 2017).

Berdasarkan permasalahan diatas, penggunaan teknik data mining dengan algoritma C 5.0 yang diimplementasikan dalam bahasa Pemrograman R diharapkan dapat memprediksi keakuratan analisa keberhasilan studi santri. dengan memantau hasil belajar santri (Haryati, Sudarsono and Suryana, 2015). Data didapatkan dari hasil *tracer study* alumni pondok pesantren beserta riwayat akademik terdahulu selama dibangku sekolah menengah atas yang akan diproses untuk mendapatkan pola *rule* yang akan menjadi landasan dalam melakukan prediksi tingkat keberhasilan studi kinerja santri.

Algoritma C 5.0 sendiri merupakan salah satu solusi pemecahan kasus yang sering digunakan pada teknik klasifikasi. Keluaran dari algoritma C 5.0 adalah berupa *tree* dan *rule based model*. Algoritma ini adalah pengembangan dari algoritma C 4.5 dan IDE (*Iterative Dichotomiser 3*) algoritma C 5.0 memiliki fitur yang lebih lengkap, lebih cepat, lebih efisien dan menghasilkan *tree* yang lebih sederhana dari C 4.5 (Kumar Mandal, 2017). Algoritma C.50 dengan masing-masing rangkaian pembagiannya, anggota himpunan hasil menjadi mirip satu dengan yang lain (Berry and Linoff, 2004). Dalam algoritma C 5.0 pemilihan atribut dilakukkan dengan menggunakan *information gain*, *gain ratio* dengan mencari nilai *entropy*. Algoritma C 5.0 mirip dengan pembangunan algoritma C4.5 kemiripan tersebut meliputi perhitungan kemunculan kejadian, perhitungan *entropy* dan *information gain*. Jika pada algoritma C 4.5 berhenti sampai

ISSN Print : 1979-7141
ISSN Online : 2541-1942

perhitungan *information gain*, maka pada algoritma C 5.0 akan melanjutkannya dengan perhitungan *gain ratio* dengan menggunakan *information gain* dan *entropy* yang telah ada. Serta algoritma C 5.0 memiliki fitur yang lebih lengkap, lebih cepat, lebih efisien dan menghasilkan *t*ree yang lebih sederhana dari C 4.5. dan membagi data berdasarkan kriteria yang dipilih untuk membuat sebuah *Decision Tree* dengan menggunakan pendekatan secara *top-down* (Wei and You, 2011).

Berdasarkan analisis yang dilakukan Johan Jansson dalam penelitiannya, algoritma C 5.0 mampu memberikan hasil yang efektif dalam mendukung suatu keputusan dengan kriteria yang dibuat secara *random.* Selain itu, alasan memilih menggunakan algoritma C 5.0 adalah mampu menghasilkan sub sistem *model base* yang dapat digunakan untuk menunjang sistem pendukung keputusan (Al-Hegami, 2007). Penelitian ini dilakukan untuk mengetahui tingkat keberhasilan studi santri berdasarkan beberapa kriteria.

2. LANDASAN TEORI

2.1. Prediksi

Digunakan untuk memperkirakan atau *forecasting* suatu kejadian sebelum kejadian – kejadian atau peristiwa tertentu terjadi. Misalnya, bagaimana Badan Meterologi Dan Geofisika (BMKG) memperkirakan tanggal tertentu bagaimana cuacanya apakah hujan, panas dan lain sebagainya. Metode yang sering digunakan salah satunya adalah *Roug se*t.

Data mining juga sama halnya dengan konsep *neural network* mengandung 2 (dua) pengelompokkan yaitu :

1) *Supervised learning* merupakan pembelajaran menggunakan guru dan biasanya ditandai dengan adanya class/label/target pada himpunan data. Adapun metode-metode yang digunakan bersifat *Supervised learning* seperti metode prediksi dan klasifikasi algoritma C 5.0, metode *roug set* dan lain-lain.
2) *Unsupervised learning* merupakan pembelajaran tanpa menggunakan guru dan biasanya ditandai pada himpunan datanya dan tidak memiliki atribut keputusan atau class/label/target. Metode-metode yang bersifat *unsupervised learning* meliputi metode *estimasi*, *clustering*, *asosiasi, regresi linier, analytical hierarchy clustering* dan lain-lain.

2.2. Algoritma C 5.0

Algoritma C 5.0 merupakan algoritma turunan dari algoritma pohon keputusan yang sebelumnya yaitu algoritma C 4.5 dan sering digunakan untuk data mining. Algoritma C 5.0 memiliki peningkatan dalam hal kecepatan memori sebesar 90% dari algoritma sebelumnya yaitu C 4.5 (Wirdhaningsih *et al.*, 2013). Dan biasanya algoritma C 5.0 ini menggunakan memori lebih rendah dapa pada algortima C 4.5 seperti contohnya pada saat *rule set* pada dataset *forest,* dimana algoritma C 4.5 menggunakan kurang lebih 3GB memori sedangkan algoritma C 5.0 kurang lebih menggunakan 200MB memori. Dari segi akurasi, algoritma C 5.0 ini memiliki tingkat kesalahan yang rendah. Algortima C 5.0 juga menghasilkan pohon keputusan yang lebih kecil dan juga *rule set* yang sedikit. Tidak seperti pada algoritma C 4.5. oleh karena itu dengan menggunakan algoritma C 5.0 memungkinkan untuk menghapus atribut yang tidak memiliki keterkaitan dengan topik penelitian secara lebih baik.

Algoritma C 5.0 menghasilkan tingkat keakuratan yang lebih tinggi dalam hal prediksi. Penggunaan algoritma C 5.0 dapat menghasilkan model prediksi dengan hasil tingkat akurasi yang lebih tinggi (Hutabarat, 2018). Algoritma C 5.0 diharapkan proses penggalian informasi lebih cepat dan optimal dengan kapasitas data yang lebih besar, sehingga kesalahan yang ditimbulkan dalam pengambilan keputusan lebih diminimalkan (Manik, Pristiwanto and Tampubolon, 2018).

Salah satu keunggulan algoritma C 5.0 adalah dapat menangani atribut kontiyu dan diskrit. Langkah pertama yang dilakukan adalah menghitung nilai *entropy* dari keseluruhan atribut, lalu selanjutnya menghitung nilai *information gain* tertinggi dari seluruh atribut sehingga didapatkan atribut yang akan digunakan sebagai akar atau *parent.* Langkah selanjutnya, percabangan pada akar untuk setiap nilainya ditentukan, kemudian setiap cabang berisi kasus yang telah dibagi. Selanjutnya, perhitungan secara berulang dilakukan untuk menentukan nilai *gain.* perhitungan

ISSN Print : 1979-7141
ISSN Online : 2541-1942

tersebut berhenti ketika semua data telah dihitung memiliki persamaan pada kelasnya. Berikut persamaan untuk mencari nilai *entropy* sebelum dilakukannya perhitungan dalam mencari *information gain* :

$$Entropy = \sum_{1-i}^{n} - P_i * Log_2 (P_i) \quad (1)$$

Keterangan :
S : Himpunan Kasus
n : Jumlah Partisi S
P_i : Properti dari S_i terhadap S

Setelah *information gain* didapat, tentukanlah *information gain* yang memiliki nilai tertinggi. Itulah yang akan menjadi akar atau *parent* pada istilah pohon keputusan. Adapun persamaan yang digunakan dapat dilihat pada persamaan berikut :

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} (\frac{|Si|}{} * Entropy(Si) \quad (2)$$

Keterangan :
S : Himpunan Kasus
A : Fitur
n : Properti Si Terhadap S
|Si| : Proporsi Si terhadap S
|S| : Jumlah Kasus Dalam S

Perhitungan *gain ratio* untuk algoritma C 5.0 akan berjalan setelah perhitungan *information gain* diatas dilakukan. Perhitungan *gain ratio* selanjutnya menggunakan persamaan dibawah ini:

$$\text{Gain Ratio} = \frac{\text{Information gain (S,A)}}{\sum_{i=1}^{n} \text{Entropy } S_i} \quad (3)$$

Dengan adanya perhitungan *gain ratio* inilah yang menjadikan pembangunan *tree* pada C 5.0 lebih ringkas dibanding *tree* pada algoritma C 4.5. sehingga menyebabkan pola tingkat keberhasilan yang dihasilkan lebih sedikit dibandingkan algoritma C 4.5.

2.3. Bahasa Pemrograman R

R (Ihaka and Gentleman, 1996) adalah implementasi *open source* S yang bebas, dikembangkan secara *kooperatif*, R adalah sebuah bahasa pemrograman statistik yang kuat dan fleksibel dan lingkungan komputasi yang efektif pada standar di antara para ahli Statistik. Bahasa pemprograman R memiliki poin kuat, dan basis pengguna yang besar di antara para ahli statistik (Fox and Andersen, 2005).

R menyediakan berbagai teknik statistika (permodelan *linier* dan non-*linier*, uji statistik klasik, analisis deret waktu, klasifikasi, klasterisasi, dan sebagainya) serta grafik. R, sebagaimana S, dirancang sebagai bahasa komputer sebenarnya, dan mengizinkan penggunanya untuk menambah fungsi tambahan dengan mendefinisikan fungsi baru. Kekuatan besar dari R yang lain adalah fasilitas grafiknya, yang menghasilkan grafik dengan kualitas publikasi yang dapat memuat simbol matematika. R memiliki format dokumentasi seperti *LaTeX*, yang digunakan untuk menyediakan dokumentasi yang lengkap, baik secara daring (dalam berbagai format) maupun secara cetakan.

R memiliki ciri khas pada bagian *syntaxnya* yaitu selalu diawali dengan “>” dan bahasa R juga memiliki beberapa keunggulan diantaranya yaitu:

ISSN Print : 1979-7141
ISSN Online : 2541-1942

1. R unggul dalam segi pengelolaan data dan juga media penyimpananya, R memiliki kelebihan lainnya yaitu, ukuran file yang telah disimpan oleh R memiliki ukuran file yang kecil.
2. R memiliki layanan dalam mengoperasikan perhitungan *array* yang lengkap
3. R juga menunjang dalam hal penelitian dibidang statistik contohnya adalah menguji statistik, menguji fungsi dalam *probabilitas* dan sebaginya
4. Perangkat lunak R menyediakan tampilan grafik yang menarik bagi *user* dan juga fleksibel
5. R diciptakan dengan fungsi yaitu *multiplatform*, yang mana multiplatform tersebut memiliki arti yaitu R dapat menyesuaikan pada berbagai sistem Operasi, tidak hanya satu jenis sistem operasi saja.

3. METODE PENELITIAN

3.1. Desain Penelitian

Desain dari penelitian ini menggunakan data primer berupa *tracer study online* menggunakan media *google form* sebagai alat untuk memberikan *form* atau soal pertanyaan secara online yang akan diinformasikan oleh pengasuh pondok pesantren kepada para responden. Responden dalam penelitian ini adalah alumni pondok pesantren di wilayah Kecamatan Jombang. Kuesioner yang di gunakan terlebih dahulu dilakukan uji *validitas* dan *reliabilitas*. Responden akan mengisi kuesioner berdasarkan atribut yang didapatkan dari literatur maupun wawancara dengan pengasuh pondok pesantren. Kemudian data responden akan diolah dengan aplikasi Rstudio menggunakan metode *deccision tree* untuk mengimplementasikan Algoritma C 5.0 pada program data mining. Metode penelitian yang digunakan dalam penerapan algoritma C 5.0 dalam memprediksi tingkat keberhasilan studi kinerja santri menggunakan metode CRISP-DM.

3.2. Pengumpulan Data

Dalam penelitian ini menggunakan pendekatan kualitatif. Tempat penelitian adalah pondok pesantren di wilayah Kecamatan Jombang. Waktu penelitian dan pengambilan data pada bulan Februari sampai September 2020. Target penelitian adalah alumni santri pondok pesantren sejumlah 300 alumni dari masing-masing pondok pesantren yang berbeda di wilayah Kecamatan Jombang. Untuk memprediksi tingkat keberhasilan studi kinerja santri ini ditujukan pada santri yang sudah menempuh pendidikan sekolah menengah atas. Dengan demikian mereka dapat di prediksi tingkat keberhasilannya dengan menghitung jumlah *class* BERHASIL dan TIDAK BERHASIL.

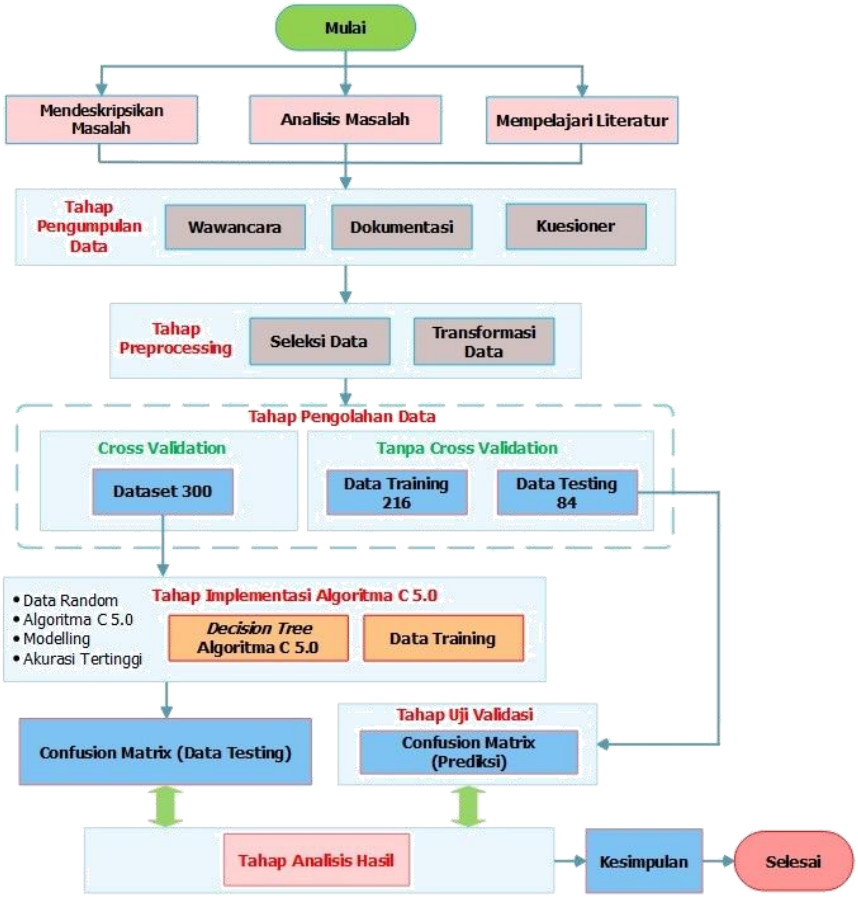
3.3. Data Preprocessing

Dalam penelitian ini dilakukan 2 teknik *preprocessing*, yaitu seleksi data dan tranformasi data. Seleksi data dilakukan secara manual dengan kreteria atribut yang dipilih meliputi hal-hal yang bersifat akademis dan erat hubungannya dengan tingkat keberhasilan studi kinerja santri. Transformasi data dilakukan untuk memperbaiki data-data yang bernilai terlalu panjang dengan menyederhanakan nilai atribut sehingga memudahkan nantinya dalam pembuatan model *decision tree* algoritma C 5.0 pada Rstudio

3.4. Alur Penelitian

Alur Penelitian yang digunakan dalam penerapan algoritma C 5.0 untuk mengetahui tingkat keberhasilan studi kinerja santri, menggunakan ilustrasi pada Gambar 1 sebagai berikut :

ISSN Print : 1979-7141
ISSN Online : 2541-1942



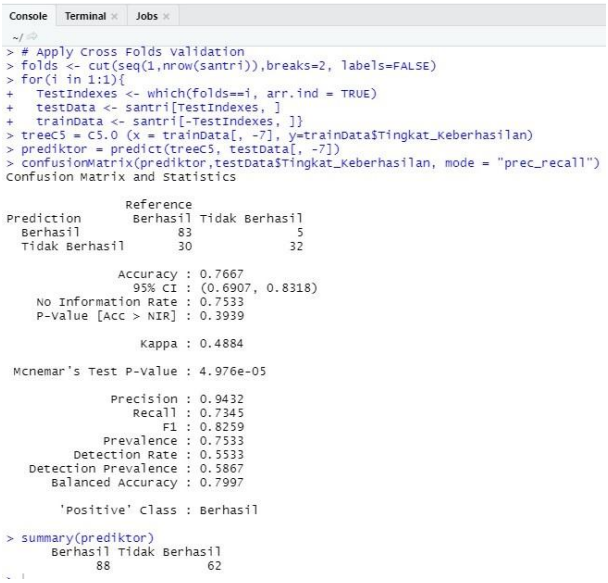
Gambar 1. Alur Penelitian

Penelitian ini dibuat untuk memprediksi faktor-faktor yang mempengaruhi tingkat keberhasilan studi kinerja santri dengan menerapkan algoritma C 5.0 dan k-*folds cross validation* serta untuk uji validasi tanpa *cross validation* menggunakan 84 Data. Hasil akhir dari penelitian ini adalah berupa model klasifikasi yang menerangkan faktor-faktor utama yang mempengaruhi dan metode k-*folds cross validation* yang menerangkan untuk melakukan evaluasi klasifikasi dengan teknik *cross validation.* Dalam pengujian ini, penulis menggunakan *2 fold, 3 fold, 6 fold, 10 fold, dan 15 fold.* Selanjutnya setelah data dikelompokkan menjadi beberapa kelompok sesuai nilai *fold,* maka langkah selanjutnya menghitung tingkat akurasi dataset tracerstudy alumni pondok pesantren.

4. HASIL DAN ANALISA

Implementasi algoritma C 5.0 dilakukan dengan bantuan perangkat lunak Rstudio dan menggunakan bahasa pemrograman R. Pembuatan model C 5.0 untuk menghasilkan nilai akurasi menggunakan Uji K *fold cross validation* dengan k = 2, 3, 6, 10 dan 15 seperti yang di tunjukkan pada gambar 2 berikut :

ISSN Print : 1979-7141
ISSN Online : 2541-1942



```
Console  Terminal  Jobs
~/
> # Apply Cross Folds Validation
> folds <- cut(seq(1,nrow(santri)),breaks=2, labels=FALSE)
> for(i in 1:1){
+   TestIndexes <- which(folds==i, arr.ind = TRUE)
+   testData <- santri[TestIndexes, ]
+   trainData <- santri[-TestIndexes, ]}
> treeC5 = C5.0 (x = trainData[, -7], y=trainData$Tingkat_Keberhasilan)
> prediktor = predict(treeC5, testData[, -7])
> confusionMatrix(prediktor,testData$Tingkat_Keberhasilan, mode = "prec_recall")
Confusion Matrix and Statistics

                Reference
Prediction       Berhasil Tidak Berhasil
  Berhasil             83              5
  Tidak Berhasil       30             32

               Accuracy : 0.7667
                 95% CI : (0.6907, 0.8318)
    No Information Rate : 0.7533
    P-Value [Acc > NIR] : 0.3939

                  Kappa : 0.4884

 Mcnemar's Test P-Value : 4.976e-05

              Precision : 0.9432
                 Recall : 0.7345
                     F1 : 0.8259
             Prevalence : 0.7533
         Detection Rate : 0.5533
   Detection Prevalence : 0.5867
      Balanced Accuracy : 0.7997

       'Positive' Class : Berhasil

> summary(prediktor)
      Berhasil Tidak Berhasil
            88             62
```

Gambar 2. Source Code K fold Cross Validation Uji Akurasi, Presisi dan Recall

Source code pada gambar 2. diulang sebanyak jumlah *fold* dengan menggunakan data santri. Hasilnya pada skenario uji K 2.1 memiliki nilai akurasi 73,45%, presisi 94,32% dan nilai recall 73,45% dan nilai TP : 83, TN : 32, FP: 5 , FN: 30, Adapun hasil keseluruhan Uji K2.1 – Uji K15.15 dapat dilihat pada tabel 1. berikut

Tabel 1. Hasil Pengukuran Akurasi, Presisi dan Recall

No	Uji Ke	TP	TN	FP	FN	Berhasil	Tidak Berhasil	Akurasi	Presisi	Recall
1	K 2.1	83	32	5	30	88	62	76,67	94,32	73,45
2	K 2.2	89	19	23	19	112	38	72	79,46	82,41
3	K 3.1	67	15	9	9	76	24	82	88,16	88,16
4	K 3.2	47	21	4	28	51	49	68	92,16	62,67
5	K 3.3	63	5	25	7	88	12	68	71,59	90
6	K 6.1	36	9	4	1	40	10	90	90	97,30
7	K 6.2	27	11	0	12	27	23	76	100	69,23
8	K 6.3	20	11	2	17	22	28	62	90,91	54,05
9	K 6.4	30	6	6	8	36	14	72	83,33	78,95
10	K 6.5	26	9	7	8	33	17	70	78,79	76,47
11	K 6.6	31	11	3	5	34	16	84	91,18	86,11
12	K 10.1	23	5	1	1	24	6	93,33	95,83	95,83
13	K 10.2	19	2	8	1	27	3	70	70,37	95
14	K 10.3	17	6	0	7	17	13	76,67	100	70,83
15	K 10.4	12	7	0	11	12	18	63,33	100	52,17
16	K 10.5	12	6	2	10	14	16	60	85,71	54,55
17	K 10.6	17	5	1	7	18	12	73,33	100	53,33
18	K 10.7	15	3	7	5	22	8	60	85,71	46,15
19	K 10.8	20	2	8	0	28	2	73,33	92,31	66,67
20	K 10.9	15	7	0	8	15	15	73,33	91,67	73,33
21	K 10.10	18	4	5	3	23	7	73,33	78,26	85,71
22	K 15.1	16	3	0	1	16	4	95	100	94,12
23	K 15.2	13	5	2	0	15	5	90	86,67	100
24	K 15.3	13	1	5	1	18	2	70	72,22	92,86
25	K 15.4	12	4	0	4	12	8	80	100	75

ISSN Print : 1979-7141
ISSN Online : 2541-1942

26	K 15.5	9	4	0	7	9	11	65	100	56,25
27	K 15.6	8	5	0	7	8	12	65	100	53,33
28	K 15.7	6	6	1	7	7	13	60	85,71	46,15
29	K 15.8	12	1	1	6	13	7	65	92,31	66,67
30	K 15.9	11	4	1	4	8	12	75	91,67	73,33
31	K 15.10	13	1	5	1	18	2	70	72,22	92,86
32	K 15.11	9	2	5	4	14	6	55	64,29	69,23
33	K 15.12	13	1	6	0	19	1	70	68,42	100
34	K 15.13	8	5	0	7	8	12	65	100	53,33
35	K 15.14	14	4	0	2	14	6	90	100	87,50
36	K 15.15	12	4	3	1	15	5	80	80	92,31
Rata-rata									97,50	92

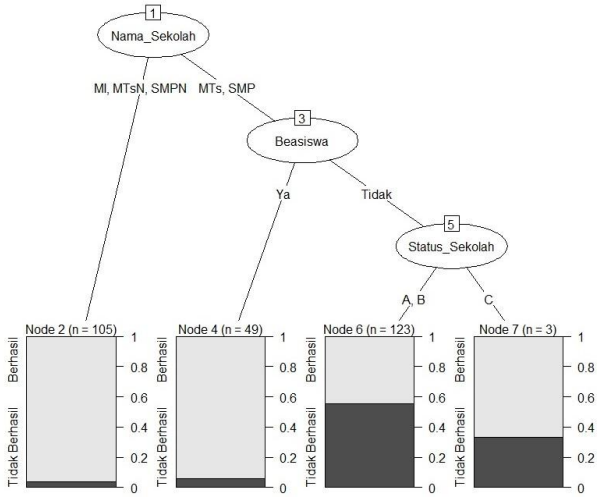
Dari tabel 1. diatas dapat diketahui bahwa nilai akurasi tertinggi diperoleh pada skenario uji K 15.1 dengan nilai 95% dan rata-rata nilai Presisi adalah bernilai 97,50% sedangkan untuk rata-rata nilai *recall* mencapai nilai 92%.

Setelah dilakukkan pengujian pada pemrograman r menggunakan 2,3,6,10 dan 15 *fold cross validation* maka hasilnya dapat dilihat pada tabel 2. Berikut ini :

Tabel 2. Tabel Hasil Akurasi, presisi dan recall dengan Pemrograman R

No	Fold	TP	TN	FP	FN	Berhasil	Tidak Berhasil	Hasil Akurasi	Hasil Presisi	Hasil Recall
1	2	83	32	5	30	88	62	76,67	94,32	73,45
2	3	67	15	9	9	76	24	82	88,16	88,16
3	6	36	9	4	1	40	10	90	90	97,30
4	10	23	5	1	1	24	6	93,33	95,83	95,83
5	15	16	3	0	1	16	4	95	100	94,12

Setelah proses hasil uji akurasi pada subbab sebelumnya, maka di hasilkan pula suatu model klasifikasi yang terbentuk dari algoritma C 5.0 dimana model tesebut direpresentasikan sebagai struktur pohon dan memuat informasi *decision tree* dari data tracer studi alumni satri yang telah diproses oleh algoritma C 5.0, berikut hasil dari *decision tree* skenario 15, 1 *folds cross validation* yang terbentuk



Gambar 3. Pohon Keputusan Akhir

ISSN Print : 1979-7141
ISSN Online : 2541-1942

Dari gambar pohon keputusan diatas dapat disimpulkan bahwa nama sekolah menjadi akar utama dalam prediksi tingkat keberhasilan studi kinerja santri selanjutnya diikuti variabel beasiswa, dan yang terakhir adalah status sekolah. Dan dari model yang sudah terbentuk menunjukkan bahwa data terlihat seimbang. Artinya hasil pembelajaran dapat melakukan prediksi dengan baik untuk class berhasil dan tidak berhasil.

4.1 Aturan – Aturan / Rule Model

Dari pohon keputusan yang terbentuk pada gambar 3. Didapat aturan-aturan / *rule model* dalam prediksi tingkat keberhasilan studi kinerja santri. Ada 4 aturan yang terbentuk, dapat dilihat sebagai berikut

1. *If* Nama_Sekolah = MI, MTsN, SMPN *then* Tingkat Keberhasilan = Berhasil
2. *If* Nama_Sekolah = MTs, SMP *And* Beasiswa = Ya *then* Tingkat Keberhasilan = Berhasil
3. *If* Beasiswa = Tidak *And* Status Sekolah = A,B *then* Tingkat Keberhasilan = Tidak Berhasil
4. *If* Status Sekolah = C *then* Tingkat Keberhasilan = Berhasil

4.2 Tahap Analisis Hasil Pengujian

Pada penelitian ini, analisa menggunakan sebuah sistem yaitu *data mining* dengan metode *Algoritma C 5.0*. Didalam proses pengekstraksian membutuhkan data tingkat keberhasilan studi santri yang didapat dari tracer alumni santri. Berikut ini adalah data sampel yang berupa tabel yang akan dilakukan proses ekstraksi sesuai dengan langkah pada metode ini.

trainData

Filter

	Nama_Sekolah	Status_Sekolah	Jumlah_Saudara	Riwayat_Sebelum_Dipesantren	Jarak_Tempuh	Beasiswa	Tingkat_Keberhasilan
21	MTsN	A	Banyak	Dipesantren	Sedang	Tidak	Berhasil
22	MTs	B	Banyak	Bersama Orangtua	Jauh	Tidak	Tidak Berhasil
23	MTs	B	Banyak	Dipesantren	Jauh	Tidak	Tidak Berhasil
24	SMPN	A	Banyak	Bersama Orangtua	Jauh	Tidak	Berhasil
25	MTsN	A	Banyak	Dipesantren	Dekat	Tidak	Berhasil
26	SMPN	B	Banyak	Bersama Orangtua	Sangat Jauh	Tidak	Berhasil
27	SMPN	A	Banyak	Bersama Orangtua	Sangat Jauh	Tidak	Berhasil
28	SMP	B	Sedikit	Dipesantren	Sangat Jauh	Ya	Berhasil
29	MTs	B	Banyak	Bersama Orangtua	Jauh	Tidak	Tidak Berhasil
30	MTsN	A	Sedikit	Dipesantren	Sangat Jauh	Ya	Berhasil

Gambar 4. Data Training

Setelah mendapatkan data training, kemudian melakukan proses perhitungan jumlah data, entropy, information gain dan gain ratio. Hasil tersebut terdapat pada tabel berikut ini

Langkah selanjutnya dilakukan perhitungan entropi total, information gain beserta gain ratio dari setiap atribut untuk menentukan *node* pertama berdasarkan tabel data sebelumnya berdasarkan ketentuan dasar entropi sebagai berikut :

ISSN Print : 1979-7141
ISSN Online : 2541-1942

Tabel 3. Perhitungan Jumlah Tingkat Keberhasilan

Node	Atribut	Nilai	Sum (nilai)	Berhasil Si	Tidak_Berhasil Si
1	Total		280	204	76
	Nama Sekolah				
		MI	23	20	3
		SMP	63	41	22
		SMPN	34	34	0
		MTs	112	62	50
		MTsN	48	47	1
	Beasiswa				
		Ya	78	75	3
		Tidak	202	129	73
	Status Sekolah				
		A	122	107	15
		B	154	94	60
		C	4	3	1

Setelah diketahui kemunculan setiap *prediktor* seperti yang terlihat pada tabel diatas, kemudian dicari nilai *entropy*. Perhitungan *entropy* pada Algoritma C 5.0, dengan Persamaan (1). Persamaan diatas berlaku pada semua atribut, termasuk atribut target, Tingkat Keberhasilan. *Entropy* pada atribut Tingkat Keberhasilan akan menjadi *entropy* total. Berikut perhitungan *entropy* dalam pemilihan *root*.

Entropy (Total) = (-(204/280) * (log2 (204/280)) + (-(76/280) * (log2 (76/280))
= 0.84350708557

Entropy (MI) = (-(20/23) * (log2 (20/23)) + (-(3/23) * (log2 (3/23))
= 0.55862937345

Entropy (SMP) = (-(41/63) * (log2 (41/63)) + (-(22/63) * (log2 (22/63))
= 0.93335726001

Entropy (SMPN) = (-(34/34) * (log2 (34/34)) + (-(0/34) * (log2 (0/34))
= NaN

Entropy (MTs) = (-(62/112) * (log2 (62/112)) + (-(50/112) * (log2 (50/112))
= 0.99170330837

Entropy (MTsN) = (-(47/48) * (log2 (47/48)) + (-(1/48) * (log2 (1/48))
= 0.14609425012

Entropy (Ya) = (-(75/78) * (log2 (75/78)) + (-(3/78) * (log2 (3/78))
= 0.23519338181

Entropy (Tidak) = (-(129/202) * (log2 (129/202)) + (-(73/202) * (log2 (73/202))
= 0.94382777607

Entropy (A) = (-(107/122) * (log2 (107/122)) + (-(15/122) * (log2 (15/122))
= 0.53778384183

Entropy (B) = (-(94/154) * (log2 (94/154)) + (-(60/154) * (log2 (60/154))
= 0.96454765891

Entropy (C) = (-(3/4) * (log2 (3/4)) + (-(1/4) * (log2 (1/4))
= 0.81127812445

ISSN Print : 1979-7141
ISSN Online : 2541-1942

Setelah perhitungan *entropy* seperti diatas, maka dilakukan perhitungan *information gain*. Perhitungan *information gain* ini menggunakan Persamaan (2). Berikut perhitungan *Information Gain* dalam penentuan *root*. Nilai pada *entropy* (S) yang dipakai adalah *Entropy* Total.

InformationGain(Nama_Sekolah,Total) = 0.84350708557 – ((23/280) * 0.55862937345) + ((63/280) * 0.93335726001) + ((43/280) * *NaN*) + ((112/280) * 0.99170330837) + ((48/280) * 0.14609425012) = 0.165888237
InformationGain(Beasiswa,Total) = 0.84350708557 – ((78/280) * 0.23519338181) + ((202/280) * 0.94382777607) = 0.097084605
InformationGain(Status_Sekolah,Total) = 0.84350708557 – ((122/280) * 0.53778384183) + ((154/280) * 0.96454765891) + ((4/280) * 0.81127812445) = 0.067096083

Perhitungan *information gain* seperti diatas yang digunakan untuk membuat *node* 1 (*root*) pada Algoritma C 5.0. Pada Algoritma C5.0, perhitungan *node* akan dilakukan berdasarkan perhitungan *gain ratio*. Untuk perhitungan ini, dapat menggunakan Persamaan (3). Berikut perhitungan *Gain Ratio* dalam penentuan *root*. Sehingga Tabel 3 diatas berubah menjadi Tabel 4 dibawah ini.

GainRatio (Nama Sekolah) = 0.165888237/ 0.558629373 + 0.93335726 + 0 + 0.991703308 + 0.14609425 = 0.196664901
GainRatio (Beaiswa) = 0.097084605/0.235193382+0.943827776 = 0.115096372
GainRatio (Status Sekolah) = 0.067096083/0.537783842+0.964547659+0.811278124 = 0.079544184

Tabel 4. Perhitungan *Entropy, Informaton Gain* dan *gain ratio*

Node	Atribut	Nilai	Sum (nilai)	Berhasil Si	Tidak_Berhasil Si	Entropy	Information Gain	Gain Ratio
	Total		279	206	73	0.843507086		
	Nama Sekolah						0.165888237	0.196664901
		MI	21	18	3	0.558629373		
		SMP	51	32	19	0.93335726		
		SMPN	32	32	0	0		
		MTs	81	47	34	0.991703308		
		MTsN	31	31	0	0.14609425		
	Beasiswa						0.097084605	0.115096372
		Ya	60	59	1	0.235193382		
		Tidak	156	101	55	0.943827776		
							0.067096083	0.079544184
	Status Sekolah							
		A	122	107	15	0.537783842		
		B	154	94	60	0.964547659		
		C	4	3	1	0.811278124		

Pada tabel diatas, nilai *gain* tertinggi terdapat pada Nama Sekolah dibandingkan dengan atribut lainnya terlihat gain tertinggi yaitu nama sekolah, nama sekolah menjadi sebuah akar karena memiliki gain tertinggi pertama.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

4.3 Validasi dan Pengujian

Dari hasil klasifikasi dan pengukuran pada data validasi dengan jumlah 84 data dengan tanpa menggunakan *k fold cross validation* diperoleh hasil sebagai berikut

Tabel 5. *Confusion Matrix*

	True Berhasil	True Tidak Berhasil
Pred. Berhasil	50	11
Pred. Tidak Berhasil	11	12

Hasil *Confusion matrix* pada tabel 5. Algoritma C 5.0 mampu mengidentifikasi sebanyak 84 data yang sesuai dengan data uji. Dari hasil data uji, 50 data bernilai berhasil dan 11 data bernilai tidak berhasil, sehingga didapat algoritma C 5.0 mampu mengidentifikasi berhasil sebanyak 50 data dan tidak berhasil sebanyak 11 data, kesalahan identifikasi sebanyak 23 data.

Analisis hasil pengujian dilakukkan dengan melakukan perhitungan secara manual dengan *confusion matrix*. Berikut ini merupakan hasil dari perhitungan *confusion matrix* pada algoritma C 5.0

$$Accuracy = \left((\frac{50 + 12}{84}) * 100\% \right) = 73{,}81\%$$

$$Precision = \left((\frac{50}{50 + 11}) * 100\% \right) = 81{,}97\%$$

$$Recall = \left((\frac{50}{50 + 11}) * 100\% \right) = 81{,}97\%$$

Dari perhitungan diatas, dapat disimpulkan bahwa hasil dari perhitungan *accuracy, precision dan recall* tersebut sama dengan hasil perhitungan yang ditampilkan pada tabel 5. berdasarkan pengujian dan analisa hasil pengujian yang dilakukan, dengan tingkat akurasi 73,81% presisi 81,97% recall 81,97% menunjukkan nilai akurasi yang masih dalam kategori baik presisi dan recall yang bernilai seimbang menyimpulkan bahwa peneliti berhasil dalam mengimplementasikan algoritma klasifikasi C 5.0 dengan baik dan akan membantu calon santri dan wali santri dalam menentukan pilihan sekolah yang tepat, apakah berhasil atau tidak.

5. KESIMPULAN

5.1 Kesimpulan

1. Proses pengumpulan data dalam penelitian ini menggunakan metode kualitatif dimana penyebaran kuesioner dilakukan dengan mengimplementasikan google form sebagai alat untuk memberikan *form* atau soal pertanyaan secara online yang akan diinformasikan oleh pengasuh pondok pesantren dan diberikan kepada alumni. Dalam kuesioner peneliti menyisipkan pertanyaan yang berhubungan erat dengan riwayat akademik terdahulu dan status sosial ketika menjadi calon santri yang digunakan sebagai parameter atribut yang paling signifikan untuk merepresentasikan tingkat keberhasilan studi santri dengan menggunakan teknik data mining dan diperoleh hasil sebagai berikut variabel nama sekolah dan variabel beasiswa adalah variabel yang mempengaruhi tingkat keberhasilan studi santri sehingga pondok pesantren dapat menjadikan landasan dalam pengaturan bagi santrinya dalam memutuskan pilihan sekolah.
2. Sedangkan dari hasil implementasi data mining dengan menggunakan Algoritma C.5.0 mampu menghasilkan *rule* guna memprediksi tingkat keberhasilan studi santri berdasarkan riwayat akademik terdahulu dan status sosial ketika masih menjadi calon santri. Pengujian decision system dengan menggunakan Aplikasi RStudio sangat dirasakan dapat mempermudah proses decission system dalam menghasilkan *rule* keputusan sebagai dasar melakukan prediksi. Dan berdasarkan hasil uji coba yang sudah dilakukan dapat diketahui bahwa dari uji validasi tanpa *cross validation* yaitu 73,81% jika dibandingkan dengan hasil yang di peroleh pada skema 15 *fold* skenario 1 yaitu 95% terjadi peningkatan 21,19%, dengan

ISSN Print : 1979-7141
ISSN Online : 2541-1942

demikian yang memliki nilai akurasi tertinggi adalah dengan menggunakan metode *cross validation*. Yang artinya algoritma ini mampu melakukan pengklasifikasian dengan baik.

5.2 Saran

1. Penelitian selanjutnya agar dapat menggunakan metode klasifikasi lain untuk menemukan tingkat akurasi, presisi dan recall lebih baik.
2. Pada penelitian ini menggunakan 300 *record*. Pada penelitian selanjutnya untuk mengestimasi akurasi yang digunakan pada sebuah algoritma akan lebih baik jika record yang digunakan lebih banyak sehingga kemungkinan akurasi akan lebih akurat dalam sebuah perhitungan algoritma.
3. Penelitian ini masih menggunakan cara manual dalam mencocokkan tingkat keberhasilan kedalam klasifikasi berhasil dan tidak berhasil dengan algoritma C 5.0 yang telah dihasilkan dari Rstudio dengan data yang ada. Pada penelitian selanjutnya akan lebih baik jika dibuatkan sistem pendukung keputusan dalam menentukan sekolah yang tepat.

ACKNOWLEDGEMENTS

Terima kasih kepada Segenap Pengasuh Pondok Pesantren di Wilayah Kecamatan Jombang yang telah memberikan izin kepada peneliti untuk menyebarkan kuesioner online. Soal pertanyaan secara online ini akan diinformasikan oleh pengasuh pondok pesantren kepada para responden. Responden dalam penelitian ini adalah alumni pondok pesantren sebanyak 300 alumni. Penelitian ini ditujukan sebagai salah satu syarat kelulusan program studi Magister Teknik Informatika pada Universitas AMIKOM Yogyakarta.

DAFTAR PUSTAKA

Al-Hegami, A. S., 2007, Classical and Incremental Classification in Data Mining Process, *IJCSNS International Journal of Computer Science and Network Security*, 7(12), pp. 179–187.

Berry, M. J. a. and Linoff, G. S., ,2004, Data mining techniques: for marketing sales and customer relationship management, *Portal.Acm.Org*.

Breiman, L. *et al.,* 1984, Classification and Regression Trees (Wadsworth Statistics/Probability), New York. *CRC Press*.

Dhofier, Z., 1994, Tradisi Pesantren, Jakarta. *VI, LP3ES*.

Fox, J. and Andersen, R., 2005, Using the R Statistical Computing Environment to Teach Social Statistics Courses, *Unpublished manuscript*.

Han, J., Kamber, M. and Pei, J., 2012, Data Mining: Concepts and Techniques, Data Mining: Concepts and Techniques. doi: 10.1016/C2009-0-61819-5.

Ihaka, R. and Gentleman, R., 1996, R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics*. doi: 10.1080/10618600.1996.10474713.

Larose, D. T., 2005, Discovering Knowledge in Data: An Introduction to Data Mining, doi: 10.1002/0471687545.

M. H. Dunham, 2003, Data Mining: Introductory and Advanced Topics. Prentice Hall, *Engineering*.

Torgo, L., 2011, Data Mining with R, Data Mining with R. doi: 10.1201/9780429292859.

Wahyuni, W. R. and Hidayati, W., 2017, Peran Sekolah dalam Membentuk Keterampilan, Wirausaha Berbasis Tauhid di SD Entrepreneur Muslim Alif-A Piyungan Bantul Yogyakarta, *MANAGERIA: Jurnal Manajemen Pendidikan Islam*. doi: 10.14421/manageria.2017.22-08.

Wei, C.-C. and You, J.-Y., 2011, C 4.5 Classifier for Solving the Problem of Water Resources Engineering, *International Journal on Advanced Science, Engineering and Information Technology*, 1(6), p. 664. doi: 10.18517/ijaseit.1.6.132.

ISSN Print : 1979-7141
ISSN Online : 2541-1942