JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER

VOL. 5. NO. 2 FEBRUARI 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864

KOMPARASI ALGORITMA *DECISION TREE, NAIVE BAYES* DAN *K-NEAREST NEIGHBOR* UNTUK MEMPREDIKSI MAHASISWA LULUS TEPAT WAKTU

Agus Budiyantara[1]; Irwansyah[2]; Egi Prengki[3]; Pandi Ahmad Pratama[4]; Ninuk Wiliani[5]

[1,2] Teknik Informatika
[1,2] STMIK Widuri, [3,4],
[1,2] www.kampuswiduri.ac.id,
agusbudiyantara@yahoo.co.id, irwanstmikwiduri@gmail.com

[3,4] Ilmu Komputer
Universitas Budi Luhur
[3,4] www.budiluhur.ac.id,
egiprengki@gmail.com, pandiap1994@gmail.com

[5]Sistem Teknologi dan Informasi
[5] Institut Teknologi dan Bisnis Bank Rakyat Indonesia (IT&B BRI)
[5]www.bri-institute.ac.id
ninukwiliani@bri-institute.ac.id

Abstract—*Private Universities (PTS) compete so tight in providing performance in producing quality graduates. In addition, the number of universities in Indonesia which counts a lot both PTN and PTS makes the higher competition between universities as well. So the university strives to improve quality and provide the best education for service recipients, namely students, where one of the problems if there are some students who are late graduating or not on time so that it becomes an obstacle to the progress of the college. Prediction of students graduating on time is needed by university management in determining preventive policies related to early prevention of Drop Out (DO) cases. This prediction aims to determine the academic factors that influence the period of study and build the best prediction model with Data Mining techniques. There are 11 attributes used for Data Mining Classification, namely NPM, Gender, Age, Department, Class, Occupation, Semester 1 Achievement Index, Semester 2 Achievement Index, Semester 3 Achievement Index, Semester 4 Achievement Index and Information as result attributes. From the results of evaluations and validations that have been carried out using the RapidMiner tools the accuracy of the Decision Tree (C4.5) method is 98.04% in the 3rd test. The accuracy of the Naïve Bayes Method is 96.00% in the 4th test. And the accuracy of the K-Nearest Neighbor Method (K-NN) of 90.00% in the second test.*

Keywords: *Prediction, Decision Tree, Naïve Bayes, K-Nearest Neighbor.*

Intisari— Perguruan Tinggi Swasta (PTS) bersaing begitu ketat dalam memberikan performanya dalam mencetak lulusan-lulusan berkualitas. Selain itu, jumlah perguruan tinggi di Indonesia yang terhitung banyak baik PTN maupun PTS membuat persaingan antar perguruan tinggi semakin tinggi pula. Maka pihak perguruan tinggi berupaya dalam meningkatkan kualitas serta memberikan pendidikan terbaik bagi penerima jasanya yaitu mahasiswa dimana salah satu masalah apabila ada beberapa mahasiswa yang terlambat lulus atau tidak tepat pada waktunya sehingga menjadi kendala untuk kemajuan dari perguruan tinggi. Prediksi mahasiswa lulus tepat waktu dibutuhkan oleh manajemen Perguruan Tinggi dalam menentukan kebijakan preventif terkait pencegahan dini kasus *Drop Out* (DO). Prediksi ini bertujuan untuk menentukan faktor akademis yang berpengaruh terhadap masa studi dan membangun model prediksi terbaik dengan teknik *Data Mining*. Atribut yang digunakan untuk Klasifikasi *Data Mining* ada 11 atribut yaitu NPM, Jenis Kelamin, Usia, Jurusan, Kelas, Pekerjaan, Indek Prestasi Semester 1, Indek Prestasi Semester 2, Indek Prestasi Semester 3, Indek Prestasi Semester 4 dan Keterangan sebagai atribut hasil. Dari hasil evaluasi dan validasi yang telah dilakukan menggunakan *tools* RapidMiner diperoleh hasil akurasi dari Metode *Decision Tree* (C4.5) sebesar 98.04% pada pengujian ke 3. akurasi Metode *Naïve Bayes* sebesar 96.00% pada pengujian ke 4. Dan akurasi Metode *K-Nearest Neighbor* (K-NN) sebesar 90.00% pada pengujian ke 2.

Kata kunci : Prediksi, *Decision Tree*, *Naïve Bayes*, *K-Nearest Neighbor*.

PENDAHULUAN

Seiring perkembangan dunia pendidikan Indonesia, Perguruan Tinggi Negeri (PTN) maupun Perguruan Tinggi Swasta (PTS) bersaing begitu ketat dalam memberikan performanya dalam mencetak lulusan-lulusan berkualitas. Selain itu, jumlah perguruan tinggi di Indonesia yang terhitung banyak baik PTN maupun PTS membuat persaingan antar perguruan tinggi semakin tinggi pula. Maka, pihak perguruan tinggi berupaya dalam meningkatkan kualitas serta memberikan pendidikan terbaik bagi penerima jasanya yaitu mahasiswa (Soni Akhmad Nulhaqim1, R. Dudy Heryadi2, Ramadhan Pancasilawan3, 2015). Dimana salah satu masalah apabila ada beberapa mahasiswa yang terlambat lulus atau tidak tepat pada waktunya sehingga menjadi kendala untuk kemajuan dari perguruan tinggi tersebut. Apabila suatu sistem dapat memperkirakan atau memprediksi mahasiswa lulus tepat waktu akan sangat mempermudah bagi pihak kampus dalam mengambil langkah-langkah pencegahan agar tidak terjadi kasus *Drop Out* (DO).

Para pembuat kurikulum dapat menggunakan hasil prediksi untuk menyelaraskan perubahan dari kurikulum dan mengevaluasi efek dari perubahan tersebut. Penasehat akademik dapat merujuk ke hasil prediksi ketika memberikan nasihat kepada para mahasiswa yang terdeteksi kemungkinan terlambat lulus sehingga tindakan pencegahan dapat diambil lebih awal (Kartini, 2017). Pemahaman kesadaran dari para pengajar, pendidikan secara personil dan manajemen akademik sehingga dapat membantu untuk memilih langkah yang tepat untuk menghasilkan penurunan tingkat *Drop Out* (D0).

Prediksi ini bertujuan untuk menentukan faktor akademis yang berpengaruh terhadap masa studi dan membangun model prediksi terbaik dengan teknik *data mining* (Kartini, 2017). Masalah menurunnya tingkat kelulusan mahasiswa tepat waktu yang dialami kampus STMIK Widuri harusnya bisa memanimalisir dan ditangani dengan metode data mining yang tepat. Karena bila tidak, akan berdampak pada penumpukan data pribadi mahasiswa, data akademik, dan menghambat kegiatan akademik serta ditakutkan akan berdampak pada menurunnya citra dan nama baik kampus. Tentu saja hal tersebut harus segera ditangani dan dicari solusi yang tepat (Sani, 2016).

Penelitian dari (Sujai, Purwanto, & H.Himawan, 2016). Dengan topik Prediksi Hasil Penjurusan Siswa Sekolah Menengah Atas Dengan Menggunakan Algoritma *Decision Tree* C.45. Dari hasil pengujian sampai dengan tahap evaluasi dihasilkan kesimpulan bahwa algoritma C4.5 mendapatkan nilai akurasi tertinggi dari penggunaan tiga model kriteria *gain_ratio*, *information_gain* dan *gini_index*, menghasilkan akurasi tertinggi pada model kriteria *gain_ratio* sebesar 96,04%. Rata-rata Nilai *AUC* di antara 0.80 – 0.90 dengan klasifikasi Baik.

Penelitian dari (Mustafa & Simpen, 2015). Dengan topik Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining Studi Kasus: Data Akademik Mahasiswa STMIK Dipanegara Makassar. Tujuan dari penelitian ini untuk melakukan prediksi terhadap kemungkian mahasiswa baru dapat menyelesaikan studi tepat waktu dengan menggunakan analisis *data mining* untuk menggali tumpukan histori data dengan menggunakan *algoritma K-Nearest Neighbor* (KNN). Aplikasi yang dihasilkan pada penelitian ini akan menggunakan berbagai atribut yang klasifikasikan dalam suatu data mining antara lain nilai Ujian Nasional (UN), Asal sekolah/ Daerah, Jenis Kelamin, Pekerjaan, Penghasilan Orang Tua, Jumlah Bersaudara, dan lain-lain sehingga dengan menerapkan anallsis KNN dapat dilakukan suatu prediksi berdasarkan kedekatan histori data yang ada dengan data yang baru, apakah mahasiswa tersebut berpeluang untuk menyelesaikan studi tepat waktu atau tidak. Dari hasil pengujian dengan menerapkan algoritma KNN dan menggunakan data sampel alumni tahun wisuda 2004 s.d. 2010 untuk kasus lama dan data alumni tahun wisuda 2011 untuk kasus baru diperoleh tingkat akurasi sebesar 83,36%.

Penelitian dari (Salmu & Solichin, 2017). Dengan topik Prediksi Tingkat Kelulusan mahasiswa Tepat Waktu Mengunakan *Naive Bayes* studi Kasus UIN Syarif Hidayatullah Jakarta. Dari hasil penelitian yang telah dilaksanakan, Atribut yang digunakan untuk memprediksi kelulusan mahasiswa tepat waktu adalah jenis seleksi, pendapatan ayah, pendidikan ibu, IP semester 1 sampai dengan 4 dan sks semester 1 sampai dengan 4. Hasil Akurasi Pengujian data yang diperoleh ialah sebesar 80,72% dari 1162 data yang digunakan untuk *data training* dan 587 data untuk *data testing*.

Berdasarkan penelitian terdahulu yang telah dilakukan diduga algoritma *Decision Tree* (C4.5) akan dapat memberikan hasil akurasi yang lebih baik dalam memprediksi mahasiswa lulus tepat waktu pada STMIK Widuri Jakarta.

BAHAN DAN METODE

Bahan

Penelitian ini metode yang digunakan yaitu metode penelitian kuantitatif. Jenis penelitian yang dilakukan pada penelitian ini adalah jenis penelitian experiment, yaitu penelitian yang dilakukan dengan cara menguji kebenaran sebuah hipotesis dengan statistik yang melibatkan penyelidikan beberapa variabel dengan menggunakan tes tertentu dan menghubungkannya dengan masalah penelitian (Noviriandini & Nurajijah, 2019).

Penelitian ini menggunakan data primer yang diperoleh dari arsip data Biro Administrasi Akademik dan Kemahasiswaan (BAAK) STMIK Widuri yaitu data mahasiswa jurusan Teknik Informatika dan Sistem Informasi tahun akademik 2010/2011, 2011/2012, 2012/2013, dan 2013/2014 berjumlah 342 data kemudian sebanyak 242 data digunakan untuk data *training* dan 100 data digunakan untuk data *testing*. Atribut yang menjadi parameter sebanyak 11 atribut yaitu NPM, Jenis Kelamin, Usia, Jurusan, Kelas, Pekerjaan, Indeks Prestasi Semester 1, Indeks Prestasi Semester 2, Indeks Prestasi Semester 3, Indeks Prestasi Semester 4 dan Keterangan sebagai atribut hasil.

Tabel 1. Contoh Sampel Dataset

NPM	JK	Usia	Jur	Kelas
12411001	Perempuan	20	SI	Sore
12411003	Perempuan	18	SI	Pagi
12411004	Laki-Laki	22	SI	Sore
12411005	Laki-Laki	21	SI	Sore
12411006	Laki-Laki	22	SI	Sore
12411007	Laki-Laki	19	SI	Pagi
12411008	Laki-Laki	23	SI	Sore
12411009	Perempuan	19	SI	Pagi
12411010	Laki-Laki	18	SI	Pagi
12411011	Perempuan	19	SI	Pagi

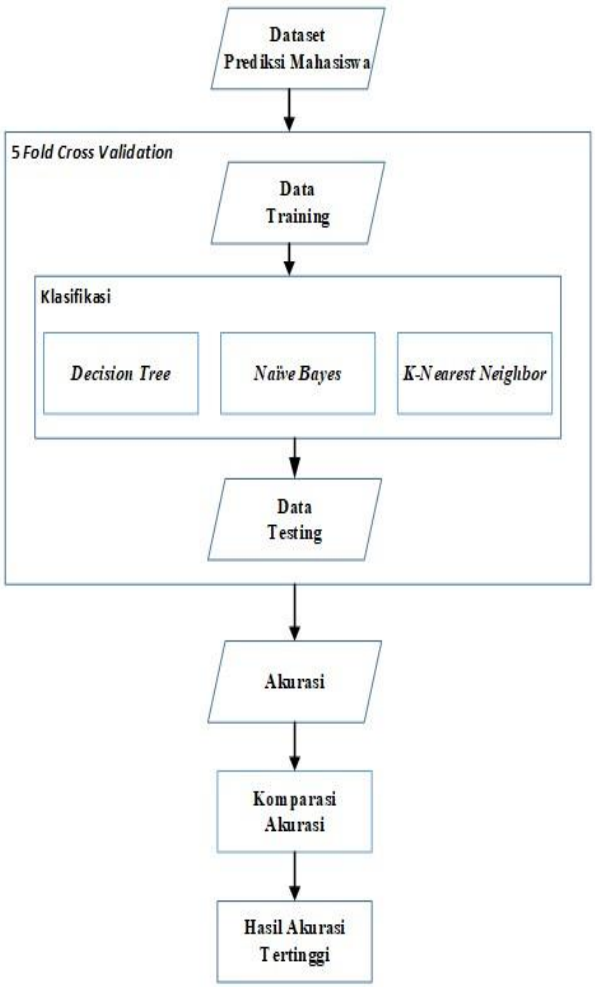
Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Tabel 2. Atribut dan Nilai Dataset

No	Nama Atribut	Nilai
1	NPM	
2	JK	- L dan P
3	Usia	- 17-25 Remaja - 26-35 Dewasa
4	Jurusan	- Sistem Informasi (SI) - Teknik Informatika (TI)
5	Kelas	- Reguler Pagi dan Sore
6	Pekerjaan	- Bekerja dan Belum Bekerja
7	IPS 1	- 0,00-1,99 Kurang Memuaskan - 2,00-2,75 Memuaskan - 2,76-3.50 Sangat Memuaskan - 3,51- 4,00 Dengan pujian
8	IPS 2	- 0,00-1,99 Kurang Memuaskan - 2,00-2,75 Memuaskan - 2,76-3.50 Sangat Memuaskan - 3,51- 4,00 Dengan pujian
9	IPS 3	- 0,00-1,99 Kurang Memuaskan - 2,00-2,75 Memuaskan - 2,76-3.50 Sangat Memuaskan - 3,51- 4,00 Dengan pujian
10	IPS 4	- 0,00-1,99 Kurang Memuaskan - 2,00-2,75 Memuaskan - 2,76-3.50 Sangat Memuaskan - 3,51- 4,00 Dengan pujian
11	KET	- Tepat - Terlambat

Sumber: (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Metode



Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

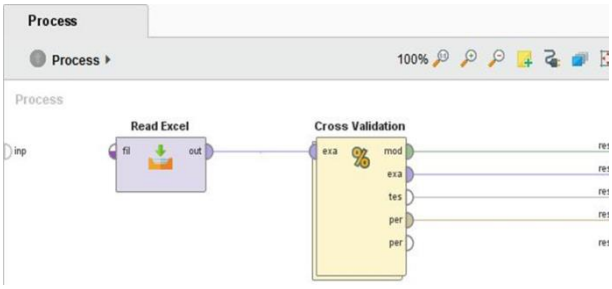
Gambar 1. Metode Komparasi *Decision Tree*, *Naïve Bayes*, *K-Nearest Neighbor*

Pada penelitian ini dilakukan dengan mengkomparasi 3 metode yaitu *Decision Tree, Naive Bayes dan K-Nearest Neighbor* seperti pada Gambar 1. Proses yang dilakukan adalah *training* dan *testing* dataset dengan menggunakan metode *Decision Tree, Naive Bayes dan K-Nearest Neighbor* untuk menghasilkan akurasi. Akurasi yang dihasilkan oleh ketiga metode tersebut kemudian dikomparasi untuk mendapatkan akurasi tertinggi (Sani, 2018).

HASIL DAN PEMBAHASAN

A. Model Proses Komparasi *Decision Tree, Naïve Bayes, dan* K-NN

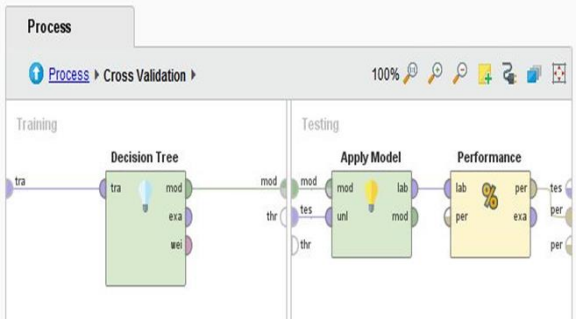
Tujuan dari penelitian adalah untuk menganalisa prediksi mahasiswa lulus tepat waktu dengan menerapkan teknik klasifikasi data mining.



Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

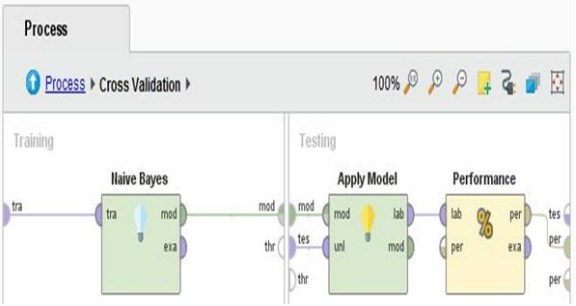
Gambar 2. Model Proses Desain *Import* Data

1. *Read Excel* : Operator ini dapat digunakan untuk memuat data dari spreadsheet Microsoft Excel.
2. *Cross Validation* : Operator yang bersarang. Ini memiliki dua subproses: subproses pelatihan dan subproses pengujian. Subproses pelatihan digunakan untuk melatih model. Model yang terlatih kemudian diterapkan dalam subproses pengujian. Kinerja model diukur selama fase Pengujian.



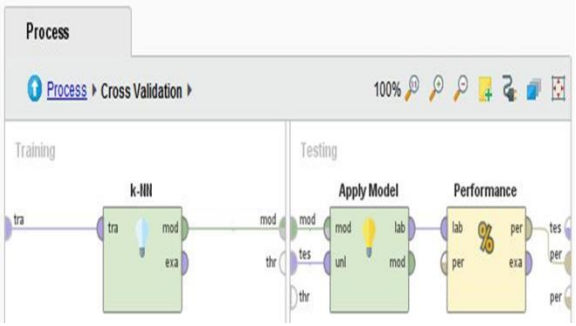
Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Gambar 3. Model Validasi C4 .5



Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Gambar 4. Model Validasi Naïve Bayes



Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Gambar 5. Model Validasi K-NN

3. *Model Validasi*: Metode klasifikasi yang digunakan dalam penelitian ini yaitu *Decision Tree* dapat dilihat pada gambar 3, *Naive Bayes* Dapat dilihat pada gambar 4, dan *K-Nearest Neighbor* dapat dilihat pada gambar 5.
4. *Apply Model* : Operator yang digunakan untuk penghubung metode *Decision Tree, Naive Bayes* dan *K-Nearest Neighbor* ke *performance*. Dapat dilihat pada gambar 3,4 dan 5.
5. *Performance* : Operator yang digunakan untuk mengukur *performance* akurasi dari model.

B. Hasil Evaluasi Model *Decision Tree* menggunakan *Cross Validation* dan *Confusion Matrix*

Table View Plot View

accuracy: 98.04% +/- 2.77% (micro average: 98.00%)

	true Terlambat	true Tepat	class precision
pred. Terlambat	40	1	97.56%
pred. Tepat	1	58	98.31%
class recall	97.56%	98.31%	

Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Gambar 6. Hasil Akurasi *Decision Tree*

Terakreditasi Sinta Peringkat 3 (Sinta 3) berdasarkan SK Dirjen Penguatan RisBang Kemenristekdikti Nomor 28/E/KPT/2019, Tanggal 26 September 2019

Dari proses evaluasi model *Decision Tree pada gambar 3* dan proses validasi terbentuk hasil matrix akurasi sebesar 98.04 % pada pengujian ke 3.

Dibawah ini merupakan perhitungan akurasi menggunakan *Confusion Matrix* dari gambar 6 diatas.

$$\text{Acuracy} = \frac{TP+TN}{TP+TN+FP+FN} + \frac{58+40}{58+40+1+1} + \frac{98}{100}$$
$$= 0{,}984 = 98{,}04\ \%$$

C. Hasil Evaluasi Model Naïve Bayes menggunakan *Cross Validation* dan *Confusion Matrix*

Table View Plot View

accuracy: 96.00% +/- 2.83% (micro average: 96.00%)

	true Terlambat	true Tepat	class precision
pred. Terlambat	38	1	97.44%
pred. Tepat	3	58	95.08%
class recall	92.68%	98.31%	

Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Gambar 7. Hasil Akurasi *Naïve Bayes*

Dari proses evaluasi model *Naïve Bayes* dan proses validasi terbentuk hasil matrix *Accuracy* sebesar 96.00% pada pengujian ke 4.

Dibawah ini merupakan perhitungan akurasi menggunakan *Confusion Matrix* dari gambar 7 diatas.

$$\text{Acuracy} = \frac{TP+TN}{TP+TN+FP+FN} + \frac{58+38}{58+38+3+1} + \frac{96}{100}$$
$$= 0{,}9600 = 96{,}0$$

D. Hasil Evaluasi Model *K-Nearest Neighbor* menggunakan *Cross Validation* dan *Confusion Matrix*

Table View Plot View

accuracy: 90.00% +/- 2.00% (micro average: 90.00%)

	true Terlambat	true Tepat	class precision
pred. Terlambat	31	0	100.00%
pred. Tepat	10	59	85.51%
class recall	75.61%	100.00%	

Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Gambar 8. Hasil Akurasi *K-Nearest Neighbors*

Dari proses evaluasi model *K-Nearest Neighbors* dan proses validasi terbentuk hasil matrix akurasi sebesar 90.00% pada pengujian ke 2.

Dibawah ini merupakan perhitungan akurasi menggunakan *Confusion Matrix* dari gambar 8 diatas.

$$\textit{Acuracy} = \frac{TP+TN}{TP+TN+FP+FN} + \frac{59+31}{58+31+10+1} + \frac{90}{100}$$
$$= 0{,}9000 = 90{,}00\ \%$$

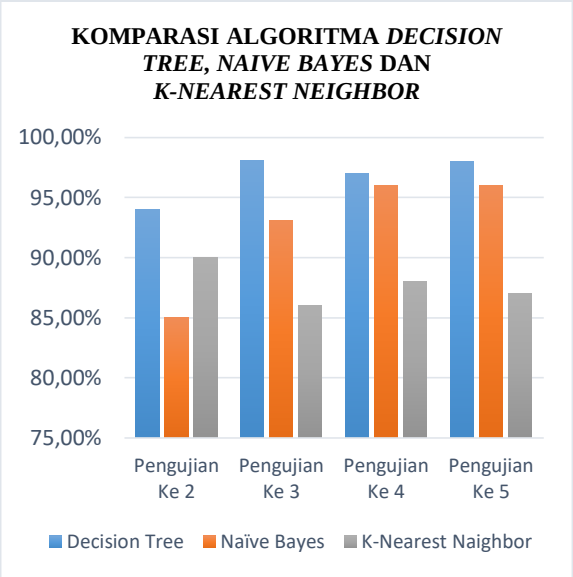
E. Hasil Pengujian Menggunakan K-5 *Cross Validation*

Tabel 3. Hasil Pengujian Akurasi

Metode	2	3	4	5
C4.5	94.00	98.04	96.94	98.00
NB	85.00	93.05	96.00	94.00
K-NN	90.00	86.04	88.00	87.00

Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Pada tabel hasil pengujian menggunakan teknik K-5 *fold cross validation* terhadap data mahasiswa dengan pengujian data mulai dari 2,3,4,dan 5.



Sumber : (Agus Budiyantara, Irwansyah, Egi Prengki, 2020)

Gambar 9. Diagram *Chart* Hasil Komparasi

Dari hasil komparasi tersebut menunjukkan *Decision Tree* memiliki tingkat akurasi yang paling tinggi. Hal tersebut menunjukkan bahwa kinerja *Decision Tree* lebih baik dibanding dengan *Naive Bayes* dan *K-Nearest Neighbor*.

KESIMPULAN

Atribut yang digunakan untuk Klasifikasi *Data Mining* terdiri atas 11 atribut yaitu NPM, Jenis Kelamin, Usia, Jurusan, Kelas, Pekerjaan, Indek Prestasi Semester 1, Indek Prestasi Semester 2, Indek Prestasi Semester 3, Indek Prestasi Semester 4 dan Keterangan sebagai atribut hasil. Dari hasil Atribut yang digunakan untuk Klasifikasi *Data Mining* terdiri atas 11 atribut yaitu NPM, Jenis Kelamin, Usia, Jurusan, Kelas, Pekerjaan, Indek Prestasi Semester 1, Indek Prestasi Semester 2, Indek Prestasi Semester 3, Indek Prestasi Semester 4 dan Keterangan sebagai atribut hasil. Dari hasil proses pengujian dengan *tools* RapidMiner Menggunakan tiga metode yang telah dilakukan. *Decision Tree* (C4.5) memperoleh hasil akurasi tertinggi sebesar 98.04% pada pengujian ke 3. Metode *Naïve Bayes* memperoleh hasil akurasi tertinggi sebesar 96.00% pada pengujian ke 4, dan Metode *K-Nearest Neighbor* (K-NN) memperoleh hasil akurasi tertinggi sebesar 90.00% pada pengujian ke 2.

REFERENSI

Agus Budiyantara, Irwansyah, Egi Prengki, P. A. P. (2020). *Komparasi Algoritma Decision Tree, Naive Bayes Dan K-Nearest Neighbor Untuk Memprediksi Mahasiswa Lulus Tepat Waktu*. Jakarta.

Kartini, D. (2017). Penerapan Data Mining dengan Algoritma Neural Network (Backpropagation) Untuk Prediksi Lama Studi Mahasiswa. *PROSIDING Seminar Nasional Sisfotek*, *3584*, 235–241.

Mustafa, M. S., & Simpen, I. W. (2015). Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining (Studi Kasus: Data Akademik Mahasiswa STMIK Dipanegara Makassar). *Creative Information Technology Journal*, *1*(4), 270. https://doi.org/10.24076/citec.2014v1i4.27

Noviriandini, A., & Nurajijah, N. (2019). Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Prestasi Siswa Sekolah Menengah Kejuruan. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *5*(1), 23–28. https://doi.org/10.33480/jitk.v5i1.607

Salmu, S., & Solichin, A. (2017). Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta. *Seminar Nasional Multidisiplin Ilmu (SENMI) 2017*, (April), 701–709.

Sani, A. (2016). *Analisa Penjualan Retail Dengan Metode Association Rule Untuk Association Rule Untuk Pengambilan Keputusan Strategis Perusahaan : 2*(June 2016), 34–50.

Sani, A. (2018). Penerapan Metode K-Means Clustering Pada Perusahaan. *Jurnal Ilmiah Teknologi Informasi*, (353), 1–7.

Soni Akhmad Nulhaqim1, R. Dudy Heryadi2, Ramadhan Pancasilawan3, M. F. 4. (2015). Peranan Perguruan Tinggi Dalam Meningkatkan Kualitas Pendidikan Di Indonesia Untuk Menghadapi Asean Community 201533. *Universitas Padjadjaran*, *6*, 198. https://doi.org/10.1017/CBO9781107415324.004

Sujai, I., Purwanto, & H.Himawan. (2016). *Prediksi Hasil Penjurusan Siswa Sekolah Menengah Atas Dengan Menggunakan Algoritma Decision Tree C.4.5*. *12*(April), 42–53.

Terakreditasi Sinta Peringkat 3 (Sinta 3) berdasarkan SK Dirjen Penguatan RisBang Kemenristekdikti Nomor 28/E/KPT/2019, Tanggal 26 September 2019