

Penggunaan Algoritma Naïve Bayes untuk Prediksi Ketepatan Waktu Lulus Mahasiswa Diploma 3

USE OF THE NAÏVE BAYES ALGORITHM TO PREDICT THE TIMELINESS OF GRADUATING DIPLOMA 3 STUDENTS

Cisde Mulyadi*¹, Lilik Sugiarto²

^{1,2} STMIK Amikom Surakarta: Jl. Veteran Notosuman Singopuran Kartasura, Sukoharjo 57164, Indonesia

¹Program Studi Manajemen Informatika STMIK Amikom Surakarta, ²Jurusan Informatika STMIK Amikom Surakarta

e-mail: *¹cisde@dosen.amikomsolo.ac.id, ²lilik@dosen.amikomsolo.ac.id

Abstrak

Salah satu masalah yang dihadapi perguruan tinggi adalah penurunan tingkat kelulusan tepat waktu. Tujuan penelitian ini adalah mengatasi permasalahan tersebut menggunakan model prediksi algoritma Naïve Bayes untuk mengetahui siapa saja mahasiswa yang diprediksi yang akan lulus tepat waktu dan yang tidak sehingga bisa dilakukan tindakan oleh program studi terkait. Enam tahapan *Cross Industry Standard Process for Data Mining* digunakan dalam penelitian eksperimen ini. Hasilnya model prediksi ini bisa memberikan tingkat akurasi sebesar 93,75%.

Kata kunci — Algoritma Naïve Bayes, Data Mining, Kelulusan Tepat Waktu, Model Prediksi

Abstract

One of the problems that colleges face is the decline in graduation rates on time. The purpose of this study is to overcome these problems using the Naïve Bayes algorithm prediction model to find out which students are predicted to graduate on time and who are not so that action can be taken by the related study program. Six stages of the Cross Industry Standard Process for Data Mining were used in this experimental research. As a result, this prediction model can provide an accuracy rate of 93.75%.

Key words — Naïve Bayes Algorithm, Data Mining, Timely Graduation, Prediction Model

1. PENDAHULUAN

Tingkat kelulusan tepat waktu adalah salah satu tolok ukur dalam menilai efektivitas pendidikan yang dilakukan program studi pada suatu perguruan tinggi, khususnya di Indonesia. Menurut Instrumen Akreditasi Program Studi (IAPS) versi 4.0 yang dirilis pada tahun 2018, sebagai contoh untuk program studi diploma tiga jika ingin mendapatkan nilai maksimal 4, maka persentase kelulusan tepat waktunya harus lebih besar atau sama dengan 50 %. Batas penilaian sebesar 50% sebenarnya mengalami penurunan dibandingkan aturan pada boring akreditasi sebelumnya yaitu sebesar 70%, namun jika tidak diperhatikan nantinya akan menjadi masalah.

Jika jumlah mahasiswa yang lulus tepat waktu bisa dijaga mendekati 100% maka akan tercipta keseimbangan antara jumlah mahasiswa masuk dengan jumlah mahasiswa yang lulus. Keseimbangan ini penting untuk dijaga dalam pengelolaan program studi di suatu perguruan tinggi. Berbagai masalah yang mungkin dihadapi antara lain penambahan jadwal mata kuliah tertentu bagi yang mengulang dan dosen membimbing tugas akhir melebihi jumlah idealnya yang diatur dalam penilaian akreditasi program studi yang pada akhirnya akan mempengaruhi penilaian aspek-aspek yang dalam akreditasi program studi seperti rasio dosen dengan mahasiswa yang tidak ideal pula. Oleh karena itu perguruan tinggi harus mampu menjaga keseimbangan keduanya yaitu salah satunya dengan cara memprediksi jumlah mahasiswa yang lulus tepat waktu. Jumlah mahasiswa yang lulus tepat waktu hasil dari prediksi ini nantinya menjadi acuan dalam penerimaan mahasiswa baru agar tercipta keseimbangan.

Permasalahan yang dihadapi perguruan tinggi yang menjadi obyek penelitian adalah persentase jumlah mahasiswa yang lulus tepat waktu menurun dari yang sebelumnya 92,7% menurun mendekati 70%. Sebagai catatan STMIK Cipta Darma Surakarta yang menjadi obyek penelitian, saat ini telah berubah nama menjadi STMIK Amikom Surakarta. Selain itu diperoleh data bahwa 5-10% mahasiswa mengalami putus studi pada saat mengambil tugas akhir (semester 6). Untuk mengatasi permasalahan tersebut salah satunya bias melakukan analisa siapa saja mahasiswa yang diprediksi yang akan lulus tepat waktu dan yang tidak tepat waktu. Untuk melakukan prediksi berdasarkan data-data masa lalu dapat dilakukan dengan teknik data mining. Dalam Data mining dapat diketahui pola data yang terjadi serta kecenderungan dari kumpulan dalam jumlah besar sehingga bermanfaat untuk dianalisis [1]. Fungsi klasifikasi data mining banyak digunakan untuk memprediksi. Adapun keluarannya berupa variabel yang nilainya polinomial [2]. Fungsi klasifikasi yang banyak digunakan antara lain Decision Trees, Naive Bayes, dan algoritma Genetik [3].

Tujuan penelitian adalah menghasilkan model prediksi kelulusan tepat waktu dari dua program studi diploma tiga yang ada pada obyek penelitian agar bisa dipakai pengelola program studi dalam meningkatkan ketepatan waktu lulus mahasiswa. Selain itu juga untuk mengurangi mahasiswa yang putus studi pada tingkat akhir. Model prediksi kelulusan tepat waktu dari penelitian ini dapat dipakai pengelola program studi dalam memprediksi kelulusan tepat waktu mahasiswa pada angkatan tahun-tahun berikutnya. Mahasiswa yang diprediksi tidak akan lulus tepat waktu bisa diberikan pendampingan khusus oleh dosen pembimbing akademik maupun pembimbing tugas akhir agar nantinya bisa lulus tepat waktu.

Studi tentang penggunaan data mining untuk melakukan prediksi ketepatan waktu lulus mahasiswa telah banyak dipublikasikan. [4] melakukan studi untuk mendapatkan klasifikasi kelulusan mahasiswa dengan Algoritma Naive Bayes menggunakan dataset Nomer Induk Mahasiswa (NIM), nama jenjang studi (Sarjana dan Diploma 3), alamat, jenis kelamin, jumlah Sistem Kredit Semester (SKS) yang telah dijalani, Indeks Prestasi Kumulatif (IPK), dan tahun lulus mahasiswa tersebut. Adapun hasilnya diperoleh nilai akurasi klasifikasi kelulusan sebesar 82,08%. Studi [5] mengembangkan aplikasi prediksi masa studi mahasiswa dengan mempertimbangkan data-data akademik mahasiswa. Algoritma Naive Bayes yang digunakan dapat memprediksi masa studi mahasiswa dengan tingkat akurasi 85.17 % pada lima semester pengujian datanya. Sedangkan [6] membuat prediksi kelulusan mahasiswa dengan Decision Tree dengan berbasis Particle Swarm Optimization dan hasil sangat akurat yaitu diperoleh akurasi sebesar 97,67%. Selanjutnya [7] melakukan perbandingan Algoritma Naive Bayes dengan C4-5 untuk menganalisis berapa lama mahasiswa lulus dan hasilnya algoritma yang bagus hasil akurasinya adalah algoritma C4-5 sebesar 84.95%, dengan selisih yang tidak terlalu banyak yaitu 1.59% dari algoritma Naive Bayes. Studi yang dilakukan [8] untuk Seleksi Calon Mahasiswa Baru dengan Algoritma Support Vector Machine (SVM) dan hasilnya diperoleh tingkat akurasi 65.00%. Metode klasifikasi Naive Bayes dipilih karena metode Naive Bayes merupakan metode probabilitas statistik yang sederhana tetapi menghasilkan hasil yang akurat dengan tingkat akurasi diatas 80%.

2. METODE PENELITIAN

Metode penelitian ini menggunakan metode eksperimen yang dikontrol sendiri. Tujuannya untuk mengetahui hubungan sebab akibat [9]. Harapannya diperoleh model yang bisa digunakan memprediksi ketepatan waktu lulus mahasiswa. Penelitian harus mengikuti kaidah tertentu yang diakui [9]. Oleh karena itu penelitian ini dilakukan dengan mengikuti tahapan dalam data mining yang memiliki enam tahapan *Cross Industry Standard Process for Data Mining* [10]. Adapun tahapan-tahapannya sebagai berikut:

1. Business Understanding

Tahap pertama ini lebih tepat disebut sebagai tahap pemahaman studi. Tahapan ini bisa

dipecah menjadi beberapa tahapan sebagai berikut.

a. Penentuan Tujuan Bisnis

Tujuan bisnis adalah tujuan studi ini. Tujuan studi ini adalah mendapatkan model prediksi ketepatan waktu lulus mahasiswa. Hasil studi ini nantinya diharapkan bisa menjadi dasar pengambilan keputusan penentuan kebijakan program studi.

b. Penilaian Situasi

Pada tahapan ini, tujuan studi perlu dipahami dan diterjemahkan ke dalam tujuan *data mining*. Berdasarkan tujuan studi yang sudah ditentukan pada tahapan sebelumnya, maka diperlukan pemahaman hal-hal apa saja yang dianggap mempengaruhi ketepatan lulus mahasiswa. Selanjutnya berdasarkan pemahaman terhadap studi-studi yang sudah dilakukan sebelumnya diperoleh data tentang atribut-atribut yang mempengaruhi ketepatan lulus mahasiswa.

c. Penentuan strategi awal *data mining*

Tahapan ini dilakukan untuk menentukan bagaimana dan darimana data tentang atribut-atribut yang mempengaruhi ketepatan lulus mahasiswa diperoleh.

2. Data *Understanding*

Tahapan berikutnya mendapatkan pemahaman yang mendalam tentang seluk beluk data mahasiswa beserta atributnya.

a. Mengumpulkan data awal

Pengumpulan data awal dari bagian terkait yang memiliki data mahasiswa beserta atributnya.

b. Mendeskripsikan data

Pada tahap ini data mahasiswa yang diperoleh dideskripsikan atribut-atributnya.

c. Mengevaluasi kualitas data

Evaluasi dilakukan dengan mencari nilai kosong/*null* pada atribut dalam *dataset* mahasiswa.

d. Memilih Atribut

Pemilihan atribut data yang tidak ada nilainya yang kosong.

3. Persiapan Data (*Data Preparation*)

Tahap ini meliputi kegiatan untuk membangun *dataset* yang akan dipakai, mulai dari data awal yang masih mentah sampai sebelum pemodelan.

a. Seleksi data

Pada tahap dilakukan seleksi terhadap atribut-atribut yang dianggap mempengaruhi kelulusan mahasiswa.

b. Preprocessing Data

Adapun yang dilakukan pada tahap ini antara lain pengurangan data yang tidak relevan serta konversi nilai yang terlalu tinggi dan rendah untuk mempermudah pembentukan model, sampai pada tahap untuk memastikan kelayakan data untuk dilakukan proses pengolahan.

c. Transformasi data

Data yang berjenis numerikal dilakukan proses inisialisasi menjadi nominal.

4. Pemodelan (*Modeling*)

Pada tahap ini teknik *data mining* akan digunakan dipilih dan ditentukan algoritmanya sesuai kebutuhan. Pada studi ini diusulkan model prediksi ketepatan waktu lulus mahasiswa menggunakan teknik data mining, yaitu dengan algoritma Naïve Bayes.

5. *Evaluation*

Evaluasi merupakan lanjutan terhadap tujuan *data mining* agar nantinya model cocok dengan sasaran dalam tahap *business understanding*.

a. Evaluation Hasil

Untuk menilai apakah hasil pemodelan sesuai dengan tujuan *data mining* dilakukan evaluasi hasil. Evaluasi dilakukan dengan metode *confusion matrix*. Nilai *recall* maupun *accuracy* diperoleh dengan menambahkan blok *performance* setelah melakukan *apply model*.

b. Review Proses

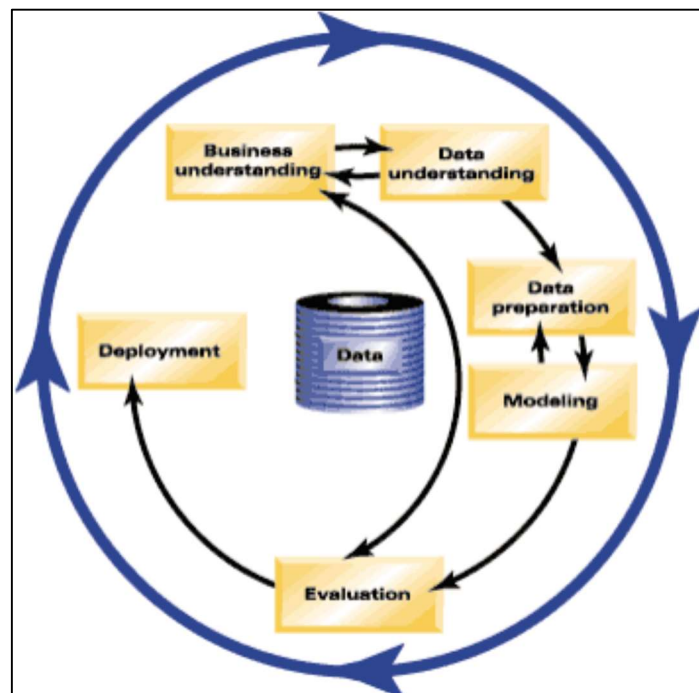
Hal ini untuk memastikan bahwa semua tahapan sudah dilaksanakan. Berdasarkan hasil pemeriksaan harus dipastikan semua tahapan telah dilakukan.

c. Menentukan Langkah Selanjutnya

Tahap ini sangat penting dilakukan karena jika hasil pemodelan belum sesuai maka harus kembali ke tahap awal sedangkan jika sudah sesuai tujuan maka bisa dilanjutkan ke tahapan deployment.

6. *Deployment*

Tahap ini merupakan tahapan pembuatan laporan hasil. Hasil analisis disusun dalam bentuk laporan yang antara lain berisi pengetahuan tentang pola pada data.



Gambar 1. Proses CRISP-DM (Larose, 2006)

3. HASIL DAN PEMBAHASAN

Adapun hasil studi ini dibahas sesuai tahapan dalam data mining yang memiliki enam fase CRISP-DM.

3.1 *Business Understanding*

Seperti yang sudah dijelaskan sebelumnya tahapan ini bisa dipecah menjadi beberapa tahapan sebagai berikut.

3.1.1 Menentukan Tujuan Bisnis

Pada tahapan ini yang dimaksud dengan tujuan bisnis adalah tujuan studi ini. Tujuan studi ini adalah mendapatkan model prediksi ketepatan waktu lulus mahasiswa. Data terhadap data alumni lulusan tahun 2017 pada salah satu program studi diploma tiga di STMIK Cipta Darma Surakarta agar nantinya dapat meningkatkan kelulusan tepat waktu mahasiswanya ke depan. Hal ini dilakukan dikarenakan banyaknya kelulusan tidak tepat waktu mahasiswanya.

3.1.2 Melakukan Penilaian Situasi

Pemahaman terhadap tujuan studi perlu diterjemahkan ke dalam tujuan *data mining*. Berdasarkan tujuan studi yang sudah ditentukan pada tahapan sebelumnya, maka diperlukan pemahaman hal-hal apa saja yang dianggap mempengaruhi ketepatan lulus mahasiswa. Selanjutnya berdasarkan pemahaman terhadap studi-studi yang sudah dilakukan sebelumnya diperoleh data tentang atribut-atribut yang mempengaruhi ketepatan lulus mahasiswa. Data Atribut-atribut mahasiswa tersebut nantinya yang akan diolah menggunakan metode *data mining*.

3.1.3 Menentukan strategi awal *data mining*

Pada tahap ditentukan bahwa data mentah bisa diperoleh dari Bagian Administrasi Akademik dan informasi Kemahasiswaan (BAIK) STMIK Cipta Darma Surakarta.

3.2 *Data Understanding*

Dataset yang diperoleh dari BAIK STMIK Cipta Darma Surakarta berupa data yang tersimpan dalam format Microsoft Excel sejumlah 317 *record*.

3.2.1 Pengumpulan data awal

Sumber data utama adalah *dataset* alumni lulusan 2017-2019 pada 2 program studi diploma tiga di STMIK Cipta Darma Surakarta dengan tipe file Microsoft Excel *Worksheet* (.xls).

3.2.2 Mendeskripsikan data

Dataset alumni tersebut terdiri dari beberapa atribut yaitu Nomer Induk Mahasiswa (NIM), nama, peminatan program studi, konsentrasi, jenis kelamin, status mahasiswa, Indeks Prestasi Kumulatif (IPK) Semester 5, nilai mata kuliah Metodologi Penelitian dan status kelulusan. IPK yang dipakai hanya sampai semester 5 dikarenakan nantinya prediksi dilakukan pada saat semester 5 berakhir, kemudian di semester 5 bisa dilakukan pemantauan.

3.2.3 Evaluasi kualitas data

Tidak terlalu banyak nilai kosong pada atribut dalam *dataset* mahasiswa sehingga dianggap kualitas data cukup baik. Nilai yang kosong juga bisa segera diperoleh dengan menanyakan kepada BAIK.

3.2.4 Pemilihan Atribut

NIM, nama, peminatan program studi, konsentrasi, jenis kelamin, status mahasiswa, IPK Semester 5, nilai mata kuliah Metodologi Penelitian dan status kelulusan merupakan atribut yang dipilih. Penggunaan atribut nilai mata kuliah Metodologi Penelitian jarang digunakan dalam model prediksi kelulusan mahasiswa. Atribut ini dianggap penting karena menentukan kompetensi mahasiswa dalam penyusunan tugas akhir.

3.3 Persiapan Data (*Data Preparation*)

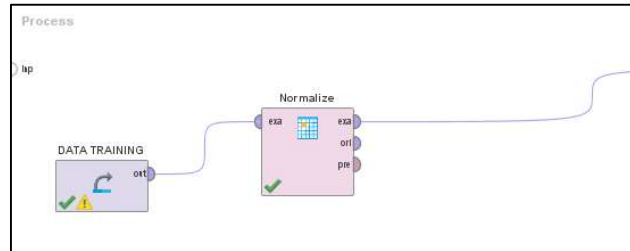
Pada Tahap Persiapan data dipecah menjadi beberapa tahapan yang lebih kecil yaitu sebagai berikut.

3.3.1 Seleksi data

Pada tahap ini akhirnya ditetapkan atribut yang digunakan tidak berubah yaitu NIM, nama, peminatan program studi, judul tugas akhir, jenis kelamin, status mahasiswa, IPK Semester 5, nilai mata kuliah Metodologi Penelitian dan status kelulusan.

3.3.2 Preprocessing Data

Seperti sudah dijelaskan sebelumnya, pada tahap ini merupakan tahap untuk memastikan data kelayakan data mahasiswa yang akan diolah sehingga perlu dilakukan secara hati-hati untuk memperoleh model yang sesuai tujuan bisnis. Pada tahap ini digunakan fitur *Normalize*.



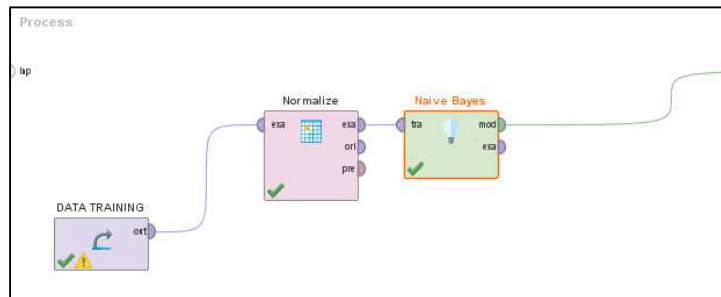
Gambar 2. Preprocessing Data dengan fitur Normalize

3.3.3 Transformasi data

Beberapa data polinomial dilakukan perubahan ke dalam bentuk binomial.

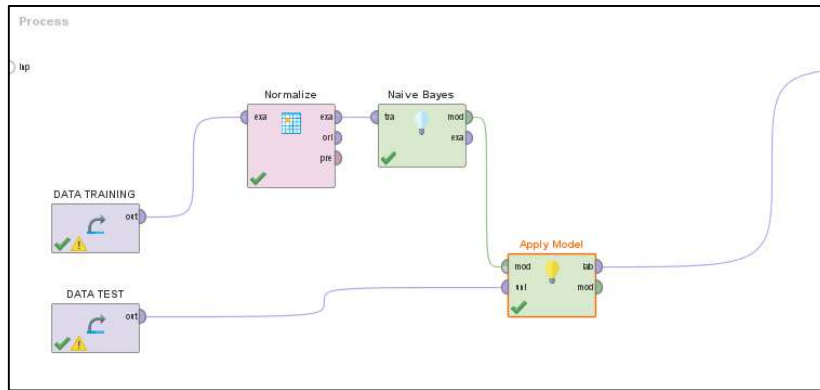
3.4 Pemodelan (*Modeling*)

Pada tahap ini teknik *data mining* mulai digunakan yaitu dengan mengusulkan model prediksi menggunakan dengan algoritma Naïve Bayes. Tahapan Pemodelan dilakukan terhadap 317 dataset alumni 2017-2019 dengan tools Rapidminer 9.8.000 seperti ditunjukkan pada gambar 2.



Gambar 3. Pemodelan dengan Naïve Bayes

Berikutnya model Naïve bayes tersebut diuji dengan menggunakan blok apply model terhadap 16 dataset alumni 2017-2019 (Data Test) yang diambil secara acak untuk model sudah bekerja seperti yang diharapkan yaitu menghasilkan prediksi. (Gambar 4.)



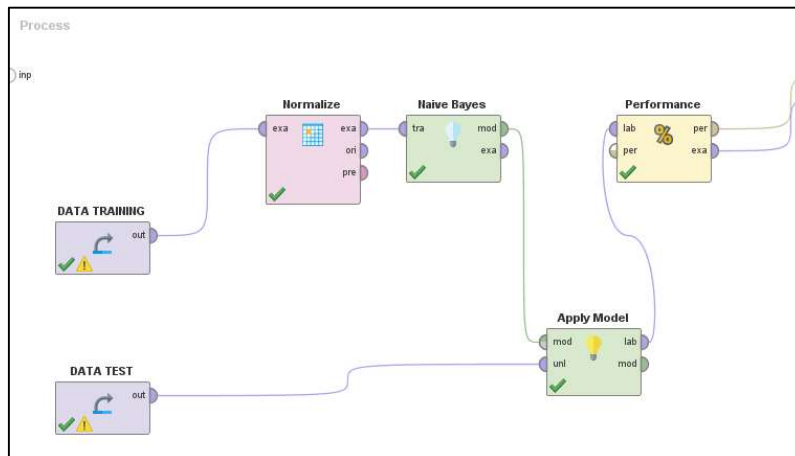
Gambar 4. Apply Model

3.5 Evaluation

Evaluasi dilakukan untuk memastikan hasil sesuai dengan sasaran yang ingin dicapai.

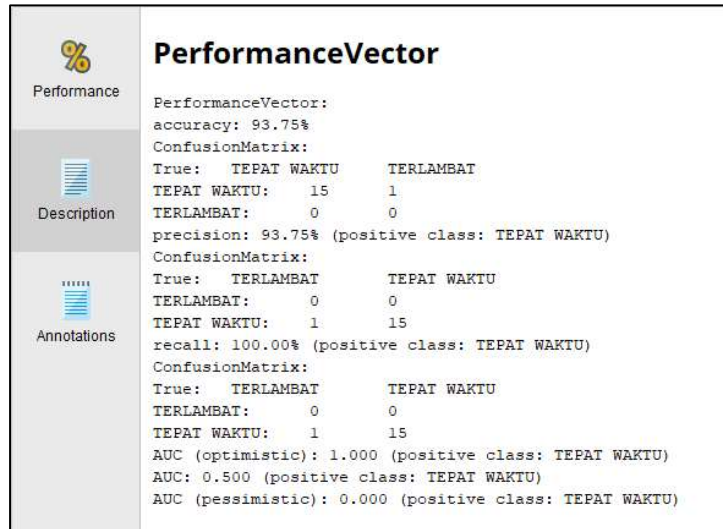
3.5.1 Evaluation Results

Evaluasi dilakukan dengan metode confusion matrix. Metode ini untuk menguji *precision*, *recall* dan juga *accuracy* pola data yang terbangun. Nilai *recall* maupun *accuracy* diperoleh dengan menambahkan blok *performance* setelah *apply model* bisa dilihat pada gambar 5.



Gambar 5. Performance

Hasilnya pengolahan dengan blok performance ditampilkan pada gambar 6. Nilai akurasi yang diperoleh sebesar 93,75%, nilai presisinya 93,75%, dan nilai recallnya 100%.



Gambar 6. Performance Vector

Berikut jika ditampilkan dalam bentuk tabel, confusion matrix dari akurasi 93,75% seperti terlihat pada gambar 7.

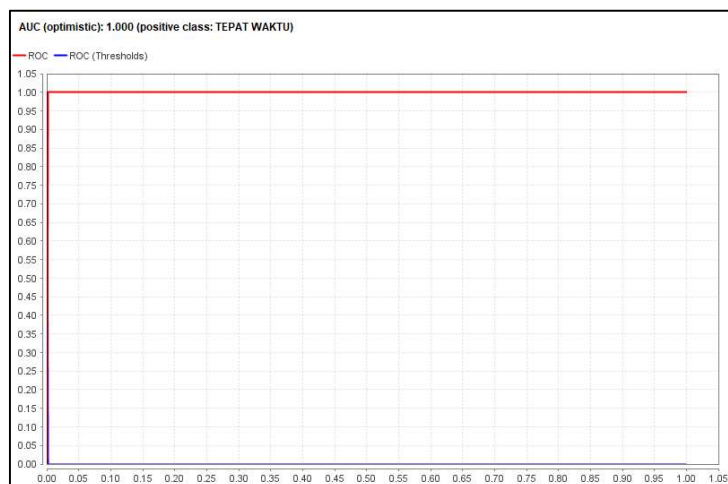
Table View Plot View

accuracy: 93.75%

	true TEPAT WAKTU	true TERLAMBAT	class precision
pred. TEPAT WAKTU	15	1	93.75%
pred. TERLAMBAT	0	0	0.00%
class recall	100.00%	0.00%	

Gambar 7. Screenshoot Hasil Akurasi

Sedangkan nilai AUC yang diperoleh sebesar 1.000 termasuk kategori *excellent classification* seperti terlihat pada gambar 8.



Gambar 8. ROC Curve

3.5.2 Review Process

Semua tahapan atau faktor penting tidak ada yang terlewatkan sehingga dilanjutkan ke tahapan berikutnya.

3.5.3 Determine Next Steps

Karena hasil pemodelan *data mining* sudah sesuai tujuan yang telah ditentukan pada tahap *business understanding* maka dilanjutkan ke tahapan *deployment*.

3.6 Deployment

3.6.1 Hasil Analisis

Studi yang dilakukan telah menghasilkan suatu informasi, dan pengetahuan baru dalam menentukan model prediksi ketepatan waktu lulus mahasiswa berdasarkan data alumni angkatan 2017-2019. Dalam rangka menilai kinerja model prediksi kelulusan mahasiswa seperti sudah dibahas sebelumnya dilakukan dengan metode confusion matrix. Berikut ditampilkan kembali hasil confusion matrix pada gambar 6. Sedangkan nilai AUC yang diperoleh sebesar 1,000 menunjukkan tingkat akurasi *excellent classification*.

4. KESIMPULAN

Algoritma Naïve Bayes mampu menghasilkan model prediksi ketepatan waktu lulus mahasiswa dengan tingkat akurasi sebesar 93,75%, presisi 93,75%, recall 100% dan nilai AUC 1,000 (*excellent classification*). Model ini dapat menghasilkan prediksi ketepatan waktu lulus mahasiswa. Data mahasiswa yang diprediksi lulus tidak tepat waktu nantinya akan menjadi masukan bagi pengelola program studi untuk memberikan penanganan khusus terhadap mahasiswa-mahasiswa tersebut agar nantinya mereka bisa lulus tepat waktu.

Studi ini masih bisa dikembangkan dengan penambahan atribut data alumni yang dianggap dapat memperbaiki prediksi sehingga hasilnya lebih akurat. Selain itu studi ini hanya menggunakan data alumni pada 2 program studi diploma 3 dan angkatan tertentu sehingga masih sangat dimungkinkan untuk menambah jumlah program studi dengan berbagai jenjang ataupun data beberapa angkatan alumni dalam satu program studi sehingga model prediksi yang diperoleh menjadi semakin baik.

UCAPAN TERIMA KASIH

Ucapan terima kasih kami sampaikan kepada Direktorat Riset dan Pengabdian Masyarakat Deputi Bidang Penguatan Riset dan Pengembangan Kementerian Riset dan Teknologi/Badan Riset dan Inovasi Nasional yang telah memberikan bantuan pendanaan dalam penelitian ini.

DAFTAR PUSTAKA

- [1] Larose, D. T., & Larose, C. D. (2015) *Data Mining and Predictive Analytics* (2nd ed.). New Jersey: John Wiley & Sons, Inc.
 - [2] Kotu, V., & Deshpande, B. (2015) *Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner*. Massachusetts: Elsevier Inc.
 - [3] Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and e-Learning*. [Online] 17(1), 118-133. Available from: doi:10.2478/eurodl-2014-0008 [Accessed 18th July 2020].
 - [4] Nugroho, Yuda Septian. (2014). *Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro. (Studi Kasus: Fakultas Ilmu Komputer Angkatan 2009)*. Skripsi. Fakultas Ilmu Komputer Universitas Dian Nuswantoro.
-

- [5] Amelia, Mongan Winny, Lumenta, Arie S. M., & Jacobus, Agustinus. (2017). Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes. *E-Journal Teknik Informatika*. Vol 11, No.1 ISSN : 2301 – 8364
- [6] Rakhman, Arif. (2017), Prediksi Ketepatan Kelulusan Mahasiswa Menggunakan Metode *Decision Tree* Berbasis *Particle Swarm Optimization*. *Smart Comp*. Volume 6, No 1.
- [7] Risqiati & Ismanto, Bambang. (2017), Analisis Komparasi Algoritma Naive Bayes Dan C4-5 untuk Waktu Kelulusan Mahasiswa. *IC-Tech*. Volume XII, No. 1
- [8] Saifudin, Aries. (2018). Metode Data Mining untuk Seleksi Calon Mahasiswa pada Penerimaan Mahasiswa Baru di Universitas Pamulang. *Jurnal Teknologi*. Volume 10, No. 1.
- [9] Dawson, C. W. (2009) *Projects in Computing and Information Systems A Student's Guide* (2nd ed.). Great Britain: Pearson Education.
- [10] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinart, T., Shearer, C. and Wirth, R. (2000), CRISP–DM Step-by-step Data Mining Guide. [Online]. Available from: www.crisp-dm.org/CRISPWP-0800.pdf [Accessed 6th August 2020).