PERBANDINGAN METODE DATA MINING MODEL KLASIFIKASI *NAIVE BAYES, DECISSION TREE* DAN K-*NEAREST NEIGHBOUR* DALAM MEMPREDIKSI KETEPATAN KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DI UNIVERSITAS PAMULANG

Ichsan Ramdhani,
Fakultas Teknik, Universitas Pamulang Tangerang Selatan
email :[1]*dosen02110@unpam.ac.id;*

ABSTRAK: Pada tahun 2019 hingga 2020 program studi teknik informatika Univesitas Pamulang telah meluluskan 1213 mahasiswa sebagai seorang sarjana komputer. Sebagian besar mahasiswa menempuh studi lebih dari waktu ideal atau mengalami "keterlambatan". Adapun tujuan dilakukannya penelitian ini adalah memberikan perbandingan dari beberapa metode *data minning* (metode naive bayes, algoritma *decission tree*, dan algoritma k-*nearest neighbour)* yang diterapkan dalam memprediksi ketepatan mahasiswa dalam menyelesaikan studinya. Pada penelitian ini terjadi penghapusan menjadi 413 data. Hal ini dilakukan karena pada kelulusan semester ganjil (800 data) semua mahasiswa terlambat. Berdasarkan data yang dikumpulkan, kelompok nilai IPK mahasiswa yang paling banyak adalah yang nilai IPKnya berkisar antara 3,01 – 3,16, yaitu 120 orang mahasiswa. Sedangkan kelompok nilai IPK yang paling sedikit adalah yang nilai IPKnya berkisar antara 3,65 – 3,80 yaitu 6 orang mahasiswa. *Data mining* dengan menggunakan model klasifikasi *naive bayes* adalah cara yang paling baik karena memiliki nilai akurasi yang lebih baik daripada *decission tree* dan algoritma k-NN. Model yang dihasilkan oleh *naive bayes* cukup baik (AUC=0,726) digunakan untuk memprediksi kelulusan mahasiswa prodi teknik informatika di Universitas Pamulang. Nilai akurasinya 75,78% dalam memprediksi ketepatan waktu lulus. Atribut kelompok nilai IPK lebih berpengaruh dalam memprediksi kelulusan tepat waktu daripada atribut jenis kelamin. Berdasarkan pohon keputusan yang diperoleh dari *decission tree,* pada kelompok nilai IPK 3,49 - 3,64, laki laki lebih memiliki peluang untuk lulus tepat waktu dibandingkan perempuan. Sedangkan pada kelompok nilai IPK 3,65 - 3,80 mahasiwi perempuan lebih berpeluang lulus tepat waktu dibandingkan mahasiswa laki-laki.

Kata Kunci: (Data mining, naive bayes, decission tree, algorima k-NN)

PENDAHULUAN

Universitas Pamulang adalah perguruan tinggi swasta di Indonesia dengan mahasiswa yang sangat banyak. Jumlah mahasiswanya kurang lebih mencapai 90.000 mahasiswa yang terdaftar. Terdapat 25 Jurusan yang berada pada 7 fakultas di Universitas Pamulang. Salah satu jurusan yang terbanyak memberikan kontribusi terhadap jumlah mahasiswa dan jumlah lulusannya adalah program studi teknik informatika. Pada setiap tahunnya jumlah mahasiswa di program studi teknik informatika mengalami peningkatan. Berdasarkan standar kelulusan mahasiswa, agar mahasiswa program studi teknik informatika dapat lulus tepat waktu maksimal masa studi berlangsung selama delapan semester atau rentang waktu 4 tahun.

Pada tahun 2019 hingga 2020 program studi teknik informatika telah meluluskan 1200 mahasiswa sebagai seorang sarjana komputer. Akan tetapi dari 1200 lulusan program studi teknik informatika lebih banyak yang menempuh studi lebih dari waktu ideal atau mengalami "keterlambatan". Ketepatan waktu lulus menjadi penting bagi program studi teknik informatika dikarenakan adanya daya tampung yang dimiliki oleh program studi teknik informatika. Daya tampung tersebut dipengaruhi oleh jumlah mahasiswa baru yang masuk dan lulusan sarjana yang telah selesai menempuh studinya.

Penggunaan basisdata oleh program studi teknik informatika dalam mendukung kegiatan belajar mengajar mahasiswa, merupakan aset berharga untuk dilakukannya analisis terhadap basisdata. Analisis tersebut dilakukan untuk mendapatkan informasi baru yang berkaitan dengan ketepatan mahasiswa untuk menyelesaikan studinya di program studi teknik informatika. Penggunaan beragam algoritma matematika yang diterapkan pada basisdata merupakan data minning untuk mencari variabel yang berpengaruh guna memprediksi ketepatan atau

keterlambatan mahasiswa dalam menyelesaikan studinya.

Adapun tujuan dilakukannya penelitian ini adalah memberikan perbandingan dari beberapa metode *data minning* (metode naive bayes, algoritma *decission tree*, dan algoritma k-*NN*) yang diterapkan dalam memprediksi ketepatan atau keterlambatan mahasiswa dalam menyelesaikan studinya. Penelitian ini dilakukan menggunakan data yang diperoleh dari kelulusan mahasiswa program studi teknik informatika universitas pamulang pada tahun 2019 sampai 2020, dengan variabel prediktor berupa Indeks Prestasi Komulatif (IPK) dan jenis kelamin dari profil lulusan.

METODE PENELITIAN

Pengumpulan Data

a. Pengumpulan Data Sekunder

Pengumpulan data pada penelitian ini dilakukan dengan memanfaatkan basisdata yang dipergunakan oleh Program Studi Teknik Informatika Universitas Pamulang. Keseluruhan data yang terkumpul sebanyak ± 1213 data, yang diperoleh dari lulusan mahasiswa Program Studi Teknik Informatika Universitas Pamulang pada tahun 2019 sampai 2020. Data yang dikumpulkan dalam penelitian ini berisikan atribut NIM, Nama, Periode Awal Kuliah, Periode Lulus, Tanggal Lulus, Indeks Prestasi Komulatif (IPK) akhir, dan Jenis Kelamin.

b. Observasi

Observasi pada penelitian ini dilakukan menggunakan observasi non partisipatif dikarenakan peneliti tidak terjun langsung dalam memperoleh data, tidak melalukan pengkuran ataupun pengamatan secara langsung. Akan tetapi yang dilakukan adalah memverifikasi data dengan melakukan pengamatan dari data yang diperoleh terhadap keadaan faktual yang ada di lingkungan Program Studi Teknik Informatika Universitas Pamulang.

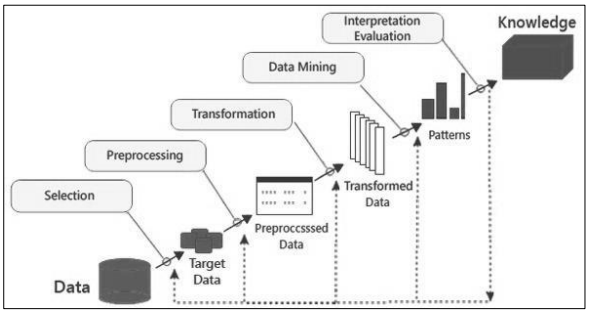
c. Studi Pustaka

Pengumpulan data menggunakan studi pustaka dilakukan dengan cara mengkaji literatur, hasil penelitian terdahulu, pemikiran dan teori yang berkaitan dengan penelitian data minning dalam menemukan ilmu pengetahuan yang baru sesuai dengan pokok kajian penelitian.

Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD), merupakan proses mencari informasi yang lebih bernilai, lebih mudah dipahami dan baru dari penyimpanan data yang besar dan kompleks. Proses KDD menafsirkan hasil yang diperoleh dari sekumpulan data dengan menggabungkan dengan ilmu lainnya. Proses KDD dimulai dengan menetapkan tujuan dan diakhiri dengan evaluasi (Tomar & Agarwal, 2013). Tahapan dari KDD dapat dilihat pada Gambar 1.

Data mining merupakan bagian dari tahapan proses KDD. Data minning menggunakan penerapan algoritma tertentu untuk mengekstrak pola dari data. Data minning menggunakan algoritma matematika untuk melakukan segmentasi terhadap data dan mengevaluasi kemungkinan dari beberapa hasil yang ditetapkan. Fungsi utama dari data minning adalah proses pengambilan pengetahuan dari volume data yang besar yang disimpan dalam repositori (Han & Kamber, 2006).



Gambar 1. Tahapan *Knowledge Discovery in Databases*

Penelitian ini merupakan jenis penelitian riset yang bertujuan menemukan pola yang dapat dipergunakan untuk memprediksi ketepatan atau keterlambatan mahasiswa dalam menyelesaikan studinya di Program Studi Teknik Informatika Universitas Pamulang. Pengambilan pengetahuan ini dilakukan terhadap basisdata yang dimiliki oleh Program Studi Teknik Informatika Universitas Pamulang, dengan memanfaatkan algoritma matematika. Tahapan dari penelitian ini adalah sebagai berikut :

a. Seleksi Data

Pada tahap ini dilakukan seleksi data lulusan dengan cara menghilangkan atribut-atribut yang tidak dibutuhkan dalam penerapan data minning. Pada data kelulusan yang telah diperoleh dari Program Studi Teknik Informatika beberapa atribut yang dihilangkan antara lain NIM, nama, periode awal kuliah, periode lulus, dan tanggal lulus.

b. Persiapan Data

Pada tahap ini dilakukan persiapan guna memenuhi kebutuhan analisis yang akan dilakukan. Data kelulusan yang diperoleh dari Program Studi Teknik Informatika kemudian dilakukan pembersihan data apabila terdapat data yang hilang, data ganda atau bersifat outlier. Selain itu pada penelitian ini terjadi penghilangan data ± 800 data, yaitu dari jumah ± 1213 data menjadi 413 data. Hal ini dilakukan melalui pertimbangan karena pada kelulusan yang diperoleh pada semester ganjil (± 800 data) kelulusannya semua terlambat. Dengan demikian variabel prediktor tidak memberikan pengaruh

apappun pada variabel target jika atribut kelulusan berada pada semester ganjil.

c. Transformasi Data

Tahapan selanjutnya setelah persiapan data, maka data dilakukan transformasi. Hal ini dimaksudkan untuk merubah bentuk dari data yang ada menjadi data yang dapat diolah menggunakan algoritma-algoritma pada data minning. Pada atribut periode masuk dan periode lulus akan diagregasi untuk menemukan lamanya kuliah. Lalu format tersebut dibuat menjadi binomial yaitu terlambat jika lamanya studi lebih dari 4 tahun, dan tepat jika lama studi kurang dari atau sama dengan 4 tahun.

Pada data jenis kelamin tidak perlu diadakan transformasi lagi dikarena jenis kelamin sudah binomial yaitu P (perempuan) dan L (laki-laki). Sedangkan untuk nilai IPK dimana nilai terkecil dalam data adalah 2,69 dan nilai tertingginya adalah 3,8, maka diubah bentuknya dari numerikal menjadi polinomial dengan 7 kategori IPK. Rincian variabel target dan variabel prediktor hasil transformasi yang akan digunakan dalam data minning ini dapat dilihat pada tabel.1

Tabel 1. Atribut Hasil Transformasi Data

Variabel Target	Kategori	
Kelulusan (Binomial)	Terlambat (> 4 tahun) Tepat (<= 4 tahun)	
Variabel Prediktor	Kategori	
Jenis Kelamin (Binomial)	P (Perempuan) L (Laki-laki)	
Indeks Prestasi Komulatif (IPK) (Polinomial)	1	2,69 - 2,84
	2	2,85 - 3,00
	3	3,01 - 3,16
	4	3,17 - 3,32
	5	3,33 - 3,48
	6	3,49 - 3,64
	7	3,65 - 3,80

d. Data Mining

Pada tahap ini dilakukan pemilihan teknik data mining yang sesuai. Untuk fungsi klasifikasi digunakan Naive Bayes, decission tree, dan algoritma k-NN. Karena klasifikasi merupakan supervised learning maka berikut ini adalah tahapan dalam model supervised learning (Larose ,2005).

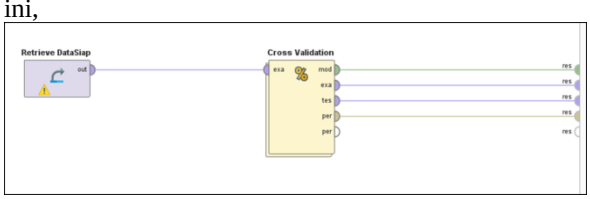
e. Evaluasi

Tahap ini digunakan untuk mengevaluasi hasil-hasil prediksi yang dihasilkan oleh ketiga model klasifikasi yaitu naive bayes, decission tree dan algoritma k-*NN* apakah mendekati klasifikasi data sebenarnya. Evaluasi dilakukan dengan menggunakan metode Confusion Matrix dan kurva ROC (Receiver Operating Characteristic). Sedang untuk indikator performansi digunakan nilai akurasi, dan AUC yang diperoleh dari ROC.

Metode Data Mining

Data mining adalah proses menelusuri pengetahuan baru, pola dan tren yang dipilah dari jumlah data yang besar yang disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola serta statistik dan teknik matematika (Larose, 2005). Terdapat beberapa algoritma atau teknik yang dapat digunakan untuk data minning. Pada penelitian ini digunakan model klasifikasi dalam data minning untuk mengambil pengetahuan baru mengenai ketepatan atau keterlambatan kelulusan mahasiswa dalam menyelesaikan studinya, yang diprediksi menggunakan nilai IPK dan jenis kelamin mahasiswa.

Pengolahan data pada penelitian ini menggunakan perangkat lunak rapidminer 9.10.011. Jumlah data yang akan diolah sebanyak 413 data. Input data yang digunakan pada rapid miner menggunakan format excel yang berisikan data yang telah melalui tahapan transformasi. Data yang menjadi inputkan terhubung pada proses cross validation sehinga terbagi atas bagian training dan testing. Teknik validasi yang digunakan untuk pada proses klasifikasi adalah k-Fold Cross Validation dengan nilai k=10. Tampilan pada proses validasi pada *rapid miner* dapat dilihat pada gambar 2 berikut ini,



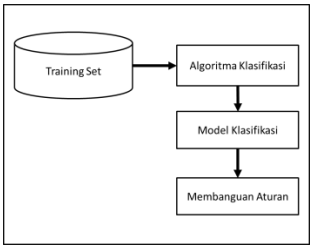
Gambar 2. Proses validasi model pada *rapid miner*

a. Klasifikasi

Klasifikasi adalah proses penemuan model yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Kamber, 2006). Proses klasifikasi dibagi menjadi dua tahap yaitu :

Tahap membangun model

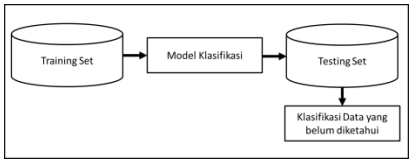
Pada langkah ini model klasifikasi dibangun berdasarkan data yang telah ditentukan kelasnya. Data sampel yang digunakan disebut sebagai data pelatihan atau data pembelajaran (training set). Proses ini disebut sebagai proses induksi yang ditunjukkan pada gambar 3.



Gambar 3. Tahapan membangun model klasifikasi

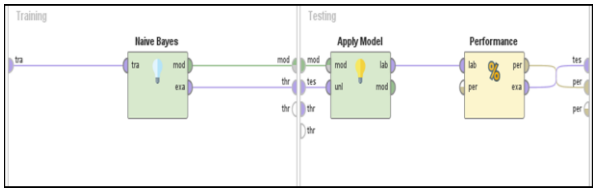
Pada tahap ini model diterapkan pada data yang belum diketahui kelasnya. Proses penerapan model klasifikasi untuk memprediksikan kelas label dari data dalam himpunan menggunakan data uji (testing set),

proses ini disebut deduksi. Proses ini dapat dilihat pada gambar. 4.



Gambar 4. Tahap penerapan model klasifikasi

Pada penelitian ini, model klasifikasi data mining yang digunakan adalah *naive bayes*, *decission tree*, dan algoritma k-*nearest neighbour*. Pada pembangunan model naive bayes di rapid miner model naive bayes dihubungkan keinput data pada area training, dengankan *apply model* dan *performance* di bagian testing. Performance ditujukan untuk mengukur kinerja sesuai dengan indikator evaluasi dari model yang telah dibangun menggunakan *naive bayes.* Perancangan model *naive bayes* untuk bagian training dan testing dapat dilihat pada gambar 2



Gambar 5. Ilustrasi rancangan model pada rapid miner

Pembangunan model menggunakan *decission tree* dan algortima k-*nearest neighbour* pada *rapid miner* dilakukan serupa dengan pembangunan model *naive bayes* pada rapid miner, yaitu dengan cara model *decission tree* dan algortima k-*nearest neighbour* dihubungkan keinput data pada area training, dengankan *apply model* dan *performance* di bagian testing. .

b. Alat Evaluasi

Klasifikasi biner merupakan model statistik dan perhitungan yang membagi kumpulan data menjadi dua kelompok yaitu positif dan negatif. Pada tabel 2 ditunjukan *confusion matrix* yang digunakan untuk menjelaskan ukuran kineja dari model klasifikasi.

Tabel 2.*Confusion Matrix*

Aktual	Prediksi	
	Positive	*Negative*
Positive	*True Positive*	*False Negative*
Negative	*False Positive*	*True Negative*

Ukuran performansi termasuk ke dalam tahapan evaluasi. Beberapa ukuran performansi untuk teknik klasifikasi yaitu akurasi, error, dan Area Under Curve (AUC). Nilai AUC diperoleh dari plot Receiver Operating Characteristics (ROC). Menurut Gorunescu (2011), nilai AUC dapat dibagi menjadi beberapa kelompok yang ditunjukan pada tabel 3.

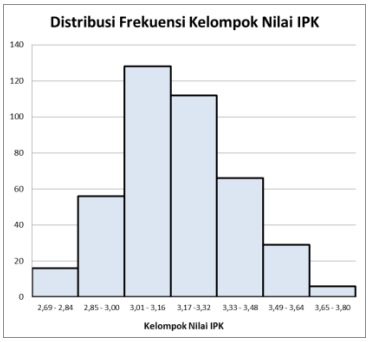
Tabel 3. Klasifikasi Nilai Area Under Curve (AUC)

No	***Nilai AUC***	**Klasifikasi**
1	0.90-1.00	Sangat baik
2	0.80-0.90	Baik
3	0.70-0.80	Cukup
4	0.60-0.70	Buruk
5	0.50-0.60	Salah

HASIL DAN PEMBAHASAN

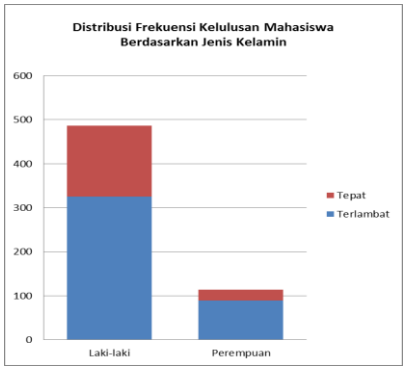
Hasil Pengolahan Data

Berdasarkan tahapan proses pembersihan data yang dilakukan, terdapat 413 data yang representatif untuk kemudian diolah menggunakan metode *data minning* yaitu *naive bayes*, *decission tree* dan algoritma k-*nearest neigbour*. Data hasil tahapan seleksi, pembersihan dan transformasi, menghasilkan variabel prediktor berupa atribut Indeks Prestasi Komulatif dan jenis kelamin mahasiswa. Berikut ini grafik distribusi frekuensi mahasiswa berdasarkan kelompok nilai IPK pada gambar 6



Gambar 6. Grafik distribusi frekuensi kelompok nilai IPK

Pada gambar, kelompok nilai IPK mahasiswa yang paling banyak adalah yang nilai IPKnya berkisar antara 3,01 – 3,16 yaitu 120 orang mahasiswa. Sedangkan kelompok nilai IPK yang paling sedikit adalah yang nilai IPKnya berkisar antara 3,65 – 3,80 yaitu 6 orang mahasiswa. Selain atribut nilai IPK, kelulusan mahasiswa prodi teknik informatika juga diprediksi menggunakan atribut jenis kelamin.



Gambar 7. Grafik distribusi frekuensi kelulusan mahasiswa berdasarkan jenis kelamin

Dari 413 data, kelulusan mahasiswa prodi teknik informatika mayoritas mengalami keterlambatan

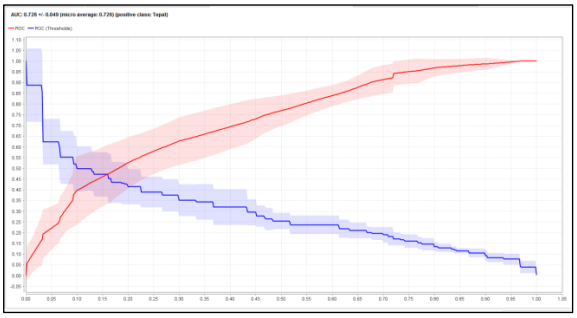
dengan masa studi lebih dari 4 tahun. Untuk mahasiswa laki-laki terdapat 67 % yang terlambat lulus, dan 33% yang lulus tepat waktu. Dan untuk mahasiswa perempuan terdapat 79 % yang mengalami keterlambatan, dan 21% dari mahasiswa perempuan lulus tepat waktu. Distribusi frekuensi kelulusan mahasiswa prodi teknik informatika berdasarkan jenis kelamin pada tahun 2020 semester genap dapat dilihat pada gambar 7.

Tahap proses *data mining* adalah bagian terpenting dari *Knowlegde Discovery in Databases* (KDD). Pada tahap ini model dibangun juga duji menggunakan data untuk kemudian dinilai seberapa tepat prediksi yang dihasilkan oleh semua model yang dibangun. Hasil dari pengembangan model menggunakan *naive bayes* diperoleh hasil uji sebagai berikut : jumlah data yang diprediksi tepat dan data sebenarnya adalah tepat sebanyak 20, jumlah data yang diprediksi tepat dan data sebenarnya adalah terlambat sebanyak 13, jumlah data yang diprediksi terlambat dan data sebenarnya adalah tepat sebanyak 87, dan jumlah data yang diprediksi terlambat dan data sebenarnya adalah terlambat sebanyak 293.

Tabel 4.*Confusion Matrix* metode *naive bayes*

Aktual	Prediksi	
	Tepat	Terlambat
Tepat	20	13
Terlambat	87	293

Berdasarkan data yang diperoleh pada tabel 4, terdapat 313 data yang sesuai antara prediksi dan data sebenarnya, dan 100 data yang tidak sesuai. Maka diperoleh nilai akurasi prediksi adalah 75,78 %.



Gambar 8. Kurva ROC untuk Naive Bayes

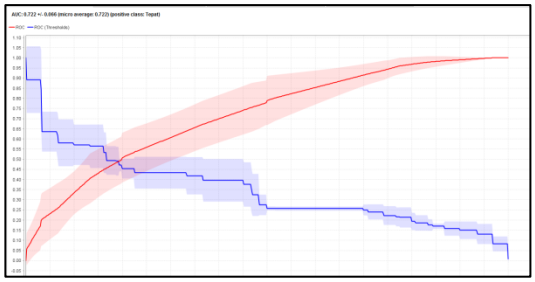
Berdasarkan plot kurva ROC untuk model yang dibangun dengan *naive bayes* diperoleh nilai AUC sebesar 0,726 yang termasuk kedalam kategori model yang cukup baik.

Hasil dari pengembangan model menggunakan *decission tree* diperoleh hasil uji sebagai berikut : jumlah data yang diprediksi tepat dan data sebenarnya adalah tepat sebanyak 23, jumlah data yang diprediksi tepat dan data sebenarnya adalah terlambat sebanyak 23, jumlah data yang diprediksi terlambat dan data sebenarnya adalah tepat sebanyak 84, dan jumlah data yang diprediksi terlambat dan data sebenarnya adalah terlambat sebanyak 283.

Tabel 5.*Confusion Matrix Decission Tree*

Aktual	Prediksi	
	Tepat	Terlambat
Tepat	23	23
Terlambat	84	283

Berdasarkan data yang diperoleh pada tabel 5, terdapat 306 data yang sesuai antara prediksi dan data sebenarnya, dan 107 data yang tidak sesuai. Maka diperoleh nilai akurasi prediksi adalah 74,09%



Gambar 9. Kurva ROC untuk *decission tree*

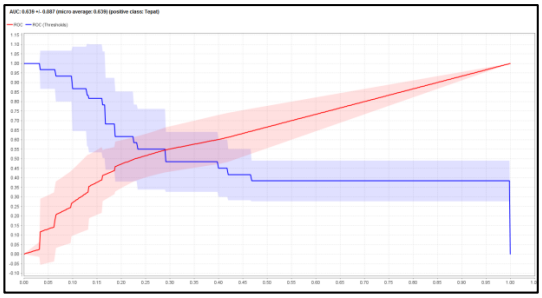
Berdasarkan plot kurva ROC untuk model yang dibangun dengan *decission tree* diperoleh nilai AUC sebesar 0,772 yang termasuk kedalam kategori model yang cukup baik. Maka pohon keputusan yang dihasilkan cukup baik untuk dipergunakan dalam memprediksi lulusan yang tepat waktu.

Hasil dari pengembangan model menggunakan algoritma k-*nearest neighbour* dengan nilai k = 3, diperoleh hasil uji sebagai berikut : jumlah data yang diprediksi tepat dan data sebenarnya adalah tepat sebanyak 42, jumlah data yang diprediksi tepat dan data sebenarnya adalah terlambat sebanyak 44, jumlah data yang diprediksi terlambat dan data sebenarnya adalah tepat sebanyak 65, dan jumlah data yang diprediksi terlambat dan data sebenarnya adalah terlambat sebanyak 262.

Tabel 6.*Confusion Matrix k-nearest neighbour (k=3)*

Aktual	Prediksi	
	Tepat	Terlambat
Tepat	42	44
Terlambat	65	262

Berdasarkan data yang diperoleh pada tabel 6, terdapat 304 data yang sesuai antara prediksi dan data sebenarnya, dan 109 data yang tidak sesuai. Maka diperoleh nilai akurasi prediksi adalah 73,64%



Gambar 10. Kurva ROC untuk algoritma k-*NN*

Berdasarkan plot kurva ROC untuk model yang dibangun dengan *decission tree* diperoleh nilai AUC sebesar 0,639 yang termasuk kedalam kategori model yang buruk. Maka model yang dibangun oleh algoritma k-*NN* tidak cukup baik untuk digunakan untuk melakukan prediksi.

Pembahasan

Ketiga model klasifikasi yang digunakan dalam *data mining* pada penelitian ini kemudian dibandingkan untuk mencari model mana yang lebih handal dalam melakukan prediksi. Terdapat 3 indikator evaluasi terhadap model prediksi yaitu *accuracy*, *error* dan *Area Under Curve* (AUC) pada plot ROC. Indikator kinerja ketiga model klasifikasi yang digunakan pada penelitian ini dapat dilihat pada tabel 7.

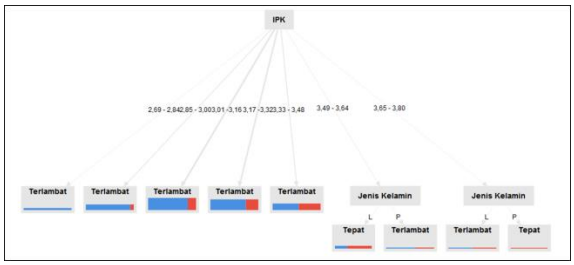
Tabel 7 Indikator kinerja model klasifikasi

Algoritma	*Accuracy*	*Error*	AUC
Naive Bayes	75,78 % ± 2.59	24,22%	0,726
Decission Tree	74,09 % ± 3.20	25,91%	0,772
k-NN, N=3	73,64 % ± 5.50	26,36%	0,639

Hasil evaluasi model klasifikasi yang digunakan pada penelitian ini, diperoleh hasil bahwa algoritma k-*nearest neighbour* merupakan model klasifikasi dengan nilai akurasi terkecil dengan nilai AUC yang masuk dalam kategori buruk (0,639). Nilai AUC tertinggi diperoleh dari model klasifikasi *decission tree* (AUC = 0,772). Sedangkan nilai AUC model klasifikasi *naive bayes* adalah 0,726. Meskipun nilai terdapat perbedaan nilai AUC pada model klasifikasi *naive bayes* dan *decission tree*, keduanya masuk dalam kategori yang sama yaitu nilai berkisar antara 0,70-0,80 yang dimaknai bahwa model yang dikembangkan cukup baik untuk melakukan prediksi.

Berdasarkan perbandingan nilai evaluasi maka model klasifikasi *naive bayes* memiliki akurasi yang paling tinggi yaitu 75,78 % ± 2.59, dengan error sebesar 24,22%, dan nilai AUC yang termasuk kedalam kategori cukup baik. Oleh karena itu ketepatan atau keterlambatan kelulusan mahasiswa program studi teknik informatika dapat diprediksi oleh model *naive bayes* yang dibangun dengan tingkat akurasi 75,78 %.

Pengembangan model klasifikasi menggunakan *decission tree* menghasilkan sebuah pohon keputusan yang akan memudahkan dalam melakukan prediksi sesuai dengan infromasi yang diperoleh dari *data mining*. Pohon keputusan dalam memprediksi tepat atau tidaknya kelulusan mahasiswa program strudi teknik informatika dapat dilihat pada gambar 11.



Gambar 11. Diagram pohon keputusan

Berdasarkan pohon keputusan yang diperoleh menggunakan model klasifikasi *decission tree* maka nilai IPK menjadi variabel prediktor yang lebih utama daibandingkan jenis kelamin. Pada kelompok nilai IPK 2,69 – 3,48 didominasi oleh mahasiswa yang lulus dengan terlambat. Untuk kelompok nilai IPK 3,49 - 3,64, laki laki lebih memiliki peluang untuk lulus tepat waktu dibandingkan perempuan. Sedangkan pada kelompok nilai IPK 3,65 - 3,80 mahasiwi perempuan lebih berpeluang lulus tepat waktu dibandingkan mahasiswa laki-laki.

KESIMPULAN

Data mining dengan menggunakan model klasifikasi *naive bayes* adalah cara yang paling baik karena memiliki nilai akurasi yang lebih baik daripada *decission tree* dan alogoritma k-NN. Model yang dihasilkan oleh *naive bayes* cukup baik (AUC=0,726) digunakan untuk memprediksi kelulusan mahasiswa prodi teknik informatika di Universitas Pamulang. Nilai akurasinya 75,78% dalam memprediksi ketepatan waktu lulus. Atribut kelompok nilai IPK lebih berpengaruh dalam memprediksi kelulusan tepat waktu daripada atribut jenis kelamin.

DAFTAR PUSTAKA

Divva, G.M.Z., dkk. 2021. *Perbandingan Metode Klasifikasi Naive Bayes, Decision Tree dan Knearest Neighbor pada Data Log Firewall*. Jakarta. SENAMIKA : e-ISBN 978-623-93343-3-8.

Gorunescu, Florin. .2011. *Data Mining Concept , Model Technique*. Springer - Verlag Berlin Heidelberg

Han, J. dan M. Kamber. 2006. *Data Mining Concepts and Techniques Second Edition*. San Francisco: Morgan Kaufmann

Kusrini dan E.T. Lutfi. 2009. *Algoritma Data Mining*. Edisi Pertama. Yogyakarta: Andi Offset.

Larose. 2005. *Discovering Knowledge in Data*. Canada: Wiley-Interscience

Suyatno. 2017. *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika

Tomar, D., & Agarwal, S. (2013). *A survey on Data Mining approaches for Healthcare*. International Journal of Bio- Science and Bio -Technology, 5 , 241-266

Widyaningsih, S. 2019. *Perbandingan Metode Data Mining Untuk Prediksi Nilai dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatik Dengan Algoritma C4.5, Naïve Bayes, KNN, dan SVM*. Jurnal Tekno Insentif | ISSN (p): 1907-4964.