PERANCANGAN APLIKASI PREDIKSI MASA STUDI MAHASISWA DENGAN METODE NAÏVE BAYES DAN C4.5

Charles Yuliansen [1] Bagus Mulyawan [2] Novario Jaya Perdana [3]

[1] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Grogol Petamburan, Jakarta Barat 11440 Indonesia
email : yuliansen92@gmail.com[1], bagus@untarfti.ac.id[2], novariojp@untarfti.ac.id[3]

ABSTRACT

In college studies each study period uses a semester system where each semester of study period will obtain a Grade Point Average (GPA) in which the GPA shows the value obtained during the study period in that semester. With the right calculation the GPA can be used as a long time determinant of a student's study period. This prediction application is made using the classification method. The data used is obtained legally from the faculty and is used to carry out the training process and test applications that have been made. For the development method use the test structure method with several tools and workmanship techniques such as flowcharts, context diagrams, and relationships between tables. The programming language used in making applications is PHP, Python, the database used is MySQL. The training and testing methods used in making the application are Naïve Bayes and C4.5.

The results of system testing show that using 237 student data obtained that the C4.5 method is always superior compared to the Naïve Bayes method. Addition of sex variables did not change the accuracy significantly.

Kata Kunci

Prediksi, Naïve Bayes, C4.5, klasifikasi, Python

1. Pendahuluan

1.1 Latar Belakang

Pendidikan tidak memiliki batas, setelah memperoleh ijazah SMA/SMK dan setarafnya maka dapat dilanjutkan menuju jenjang yang lebih tinggi, yaitu jenjang perkuliahan. Dalam pelaksanaannya kegiatan perkuliahaan merupakah salah satu kegiatan akademik utama yang dapat membantu individu dalam menentukan karir di masa mendatang. Penggunaan teknologi dapat dipergunakan dalam melakukan prediksi terhadap berbagai kasus. Prediksi dilakukan berdasarkan perhitungan dengan berbagai metode.

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (Decision Tree). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variable target. pada cabang memiliki kelas yang sama.

Naive Bayes adalah suatu klasifikasi berpeluang sederhana berdasarkan aplikasi teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen). Dalam hal ini, diasumsikan bahwa kehadiran atau ketiadaan dari suatu kejadian tertentu dari suatu kelompok tidak berhubungan dengan kehadiran atau ketiadaan dari kejadian lainnya.[1]

1.2 Rumusan Masalah

Sistem yang dirancang adalah Perancangan aplikasi perbandingan metode Naive bayes dan C4.5 untuk prediksi kelulusan mahasiswa, atribut yang akan digunakan adalah jenis kelamin, perolehan IPK peserta didik perguruan tinggi pada semester 1 (satu) hingga semester 5 (lima). Data yang digunakan dalam proses pelatihan dan pengujian program diperoleh meminta izin dari pihak yang terkait, data yang diperoleh berdasarkan mahasiswa yang telah lulus pada fakultas dengan jenjang yang sama. Kelas kelulusan yang digunakan dalam program adalah 8 semester, 9 semester, 10 semester, 11 semester dan 12 semester. Sistem aplikasi ini dirancang dengan menggunakan bahasa pemograman Python, PHP, dijalankan dalam sistem operasi berbasis Windows 10, dan berbentuk aplikasi website.

1.3 Batasan Masalah

Program aplikasi prediksi kelulusan ini memiliki beberapa batasan yaitu sebagai berikut:

1. Data yang digunakan dalam sistem pelatihan dan pengujian adalah data mahasiswa yang telah lulus menempuh studi perguruan tinggi.
2. Data yang digunakan dalam sistem pelatihan dan pengujian adalah data mahasiswa yang menjalani

perkuliahan tanpa melakukan cuti selama masa perkuliahan berlangsung.

3. Data yang digunakan dalam sistem pelatihan dan pengujian didapatkan dengan cara diperoleh dari pihak yang terkait dan data tersebut diizinkan untuk dipergunakan hanya untuk kepentingan skripsi individu oleh pihak yang terkait.

1.4 Kegunaan

Kegunaan pembuatan dan perancangan aplikasi prediksi masa studi mahasiswa dengan metode Naïve Bayes dan C4.5 adalah:

1. Membangun sistem yang dapat memprediksi kelulusan mahasiswa berdasarkan perolehan IPK dan jenis kelamin dengan tepat.
2. Menganalisis tingkat keberhasilan prediksi dengan metode *Naïve Bayes* dan *C4.5*
3. Membandingkan metode terbaik untuk melakukan prediksi masa studi mahasiswa.
4. Membantu menambah khasanah ilmu pengetahuan mengenai prediksi data dan semua metode yang digunakan di dalamnya serta memberi pondasi pembangunan sistem sejenis yang lebih efektif dan efisien di masa mendatang.

2. Landasan Teori

2.1 Prediksi

Prediksi adalah suatu proses memperkirakan secara sistematis tentang sesuatu yang paling mungkin terjadi di masa depan berdasarkan informasi masa lalu dan sekarang yang dimiliki, agar kesalahannya (selisih antara sesuatu yang terjadi dengan hasil perkiraan) dapat diperkecil dalam memperkirakan kejadian. prediksi tidak harus memberikan jawaban secara pasti kejadian yang akan terjadi, melainkan berusaha untuk mencari jawaban sedekat mungkin yang akan terjadi. [2]

2.2 Masa Studi

Masa Studi adalah masa studi terjadwal yang harus ditempuh oleh mahasiswa sesuai dengan rentang waktu yang dipersyaratkan. Tiap program yang dimiliki suatu perguruan tinggi memiliki batas maksimum dan minimum masa studi.[3]

2.3 Batas waktu Studi

Batas Waktu Studi adalah batas waktu maksimal yang diperkenankan untuk mahasiswa menyelesaikan studi. Apabila mahasiswa melampaui batas waktu studi yang ditentukan program pendidikan yang diambil, maka akan ada konsekuensi yang diterima.

2.4 IPK

IPK adalah mekanisme penilaian keseluruhan prestasi terhadap mahasiswa dalam istem perkuliahan selama masa kuliah. IPK singkatan dari Index Prestasi Kumulatif. Merupakan nilai kumulatif dari IPK (Index Prestasi). Sedangkan IPK merupakan prestasi mahasiswa per semester.[4]

2.5 PHP

PHP: *Hypertext Preprocessor* atau bahasa pemograman PHP yang biasa kita kenal adalah sebuah bahasa script sisi server atau *server-side* untuk pengembangan *website*. Dan, bahasa ini juga dapat digunakan untuk keperluan umum. Bahasa pemograman PHP ini dibuat oleh Rasmus Lerdorf pada tahun 1994.[5]

2.6 MySQL

MySQL adalah sebuah *open-source relational database management system* (RDBMS). Nama *MySQL* merupakan kombinasi dari "*My*" dan "*SQL*". Dimana, "*My*" merupakan nama anak perempuan dari co-founder "Michael Widenius" dan "*SQL"* adalah singkatan dari *Structured Query Language.*

2.7 Python

MySQL adalah sebuah *open-source relational database management system* (RDBMS). Nama *MySQL* merupakan kombinasi dari "*My*" dan "*SQL*". Dimana, "*My*" merupakan nama anak perempuan dari co-founder "Michael Widenius" dan "*SQL"* adalah singkatan dari *Structured Query Language.*[6]

2.8 Confusion Matrix

Pengukuran terhadap kinerja suatu sistem klasifikasi merupakan hal yang penting. Kinerja sistem klasifikasi menggambarkan seberapa baik sistem dalam mengklasifikasikan data. *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi.[7]

2.9 Naïve Bayes

Naive bayes adalah suatu klasifikasi berpeluang sederhana berdasarkan aplikasi teorema Bayes dengan asumsi antar variabel penjelas saling bebas (independen). Dalam hal ini, diasumsikan bahwa kehadiran atau ketiadaan dari suatu kejadian tertentu dari suatu kelompok tidak berhubungan dengan kehadiran atau ketiadaan dari kejadian lainnya. Naive Bayes dapat digunakan untuk berbagai macam keperluan antara lain untuk klasifikasi dokumen, deteksi spam atau filtering spam, dan masalah klasifikasi lainnya.

Selama proses pelatihan harus dilakukan pembelajaran probabilitas akhir P(Y|X) pada model untuk setiap kombinasi X dan Y berdasarkan informasi yang didapat dari data latih. Dengan membangun model tersebut, suatu data uji X' dapat diklasifikasikan dengan mencari nilai Y' dengan memaksimalkan nilai P(X'|Y') yang didapat.
Formulasi Naïve Bayes untuk klasifikasi adalah [8]:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(Xi|Y)}{P(X)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

2.10 C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (Decision Tree). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variable target. Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain: ID3, CART, dan C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3, Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan rule. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah meliputi, pilih atribut sebagai akar, buat cabang untuk tiap-tiap nilai, bagi kasus dalam cabang, ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *Gain* digunakan rumus berikut[9]:

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

Keterangan:
S : Himpunan kasus
A : Atribut
N : Jumlah partisi atribut A
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah kasus dalam S

Setelah mendapatkan nilai *Gain*, ada satu hal lagi yang perlu dilakukan perhitungan yaitu mencari nilai *Entropy*, *Entropy* digunakan untuk menentukan seberapa informatif sebuah input atribut untuk menghasilkan *output* atribut. Rumus dasar dari *Entropy* tersebut adalah sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} -p_i * log_2 p_i$$

S : Himpunan kasus
N : Jumlah partisi S
p_i : Proporsi dari Si terhadap S.

3. Hasil Pengujian

3.1 Proporsi data

Pengujian terhadap data dilakukan terhadap hasil percobaan. Pengujian ini dilakukan untuk mengetahui hasil dari aplikasi menganalisis prediksi masa studi yang telah dibuat. Data yang digunakan adalah data dari mahasiswa angkatan 2010-2012. Data IPK yang digunakan adalah data semester 1 hingga semester 5 dan jenis kelamin. Sistem pengujian data dilakukan dengan 3 skenario. Skenario tersebut berupa pembagian data 50:50, 70:30, 90:10. Data yang digunakan adalah data mahasiswa yang tidak melakukan cuti selama masa perkuliahan.

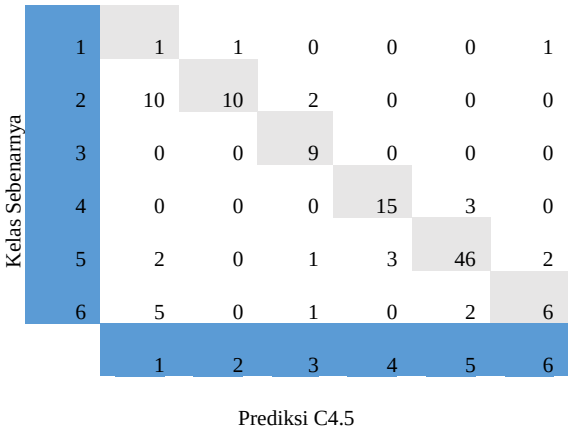
Tabel 1 proporsi data

Proporsi	Akurasi Metode	
	C4.5	Naïve Bayes
50:50	0.79831	0. 70588
70:30	0.90277	0.77777
90:10	0.875	0.70833

Pada percobaan setiap proporsi diperoleh hasil akurasi bahwa metode C4.5 selalu lebih unggul dibanding metode Naïve Bayes. Nilai akurasi yang ditampilkan pada tabel 1 adalah akurasi maksimum dari 5 skenario yang dilakukan pada tiap proporsi data.

3.2 Pengujian Evaluasi Confusion Matrix

Pengujian dilakukan dengan 3 proporsi data pertama adalah 50:50, proporsi kedua adalah 70:30 dan proporsi ketiga adalah 90:10.

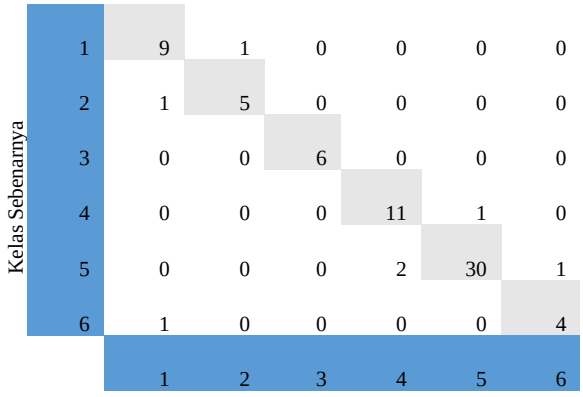


Gambar 1 Confusion Matrix 50:50

Keterangan:
1 = Lulus semester 7
2 = Lulus semester 11

3 = Lulus semester 12
4 = Lulus Semester 10
5 = Lulus Semester 8
6 = Lulus semester 9

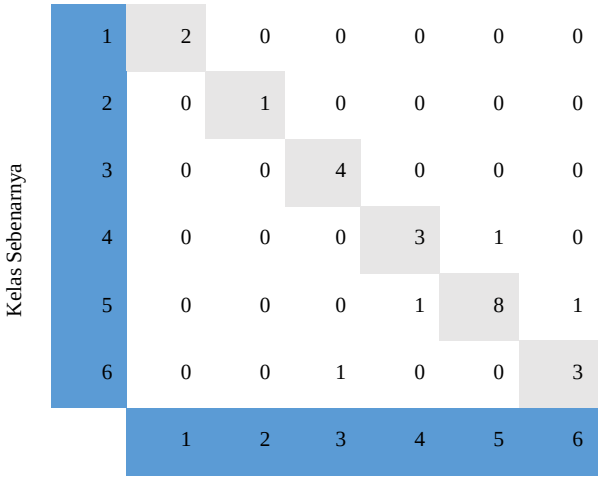
Dari hasil pengujian proporsi 50:50, dapat dilakukan pengamatan bahwa kesalahan paling umum dilakukan oleh sistem ketika melakukan prediksi dan menampilkan lulus pada semester 7, dan sistem melakukan prediksi terbaik pada Lulus semester 8. Dimana dari 51 data Sistem memprediksi 46 data.



Gambar 2 Confusion Matrix 70:30

Keterangan:
1 = Lulus semester 10
2 = Lulus semester 11
3 = Lulus semester 12
4 = Lulus Semester 7
5 = Lulus Semester 8
6 = Lulus semester 9

Dari hasil pengujian proporsi 70:30, dapat dilakukan pengamatan bahwa kesalahan paling umum dilakukan oleh sistem ketika melakukan prediksi dan menampilkan lulus pada semester 10, dan sistem melakukan prediksi terbaik pada Lulus semester 12. Dimana dari 6 data Sistem memprediksi 6 data.



Gambar 3 Confusion Matrix 90:10

Keterangan:
1 = Lulus semester 10
2 = Lulus semester 11
3 = Lulus semester 12
4 = Lulus Semester 7
5 = Lulus Semester 8
6 = Lulus semester 9

Dari hasil pengujian proporsi 90:10, dapat dilakukan pengamatan bahwa kesalahan paling umum dilakukan oleh sistem ketika melakukan prediksi dan menampilkan lulus pada semester 7 dan lulus pada semester 9, dan sistem melakukan prediksi terbaik pada Lulus semester 10 dan lulus pada semester 11. Dimana dari 6 data Sistem memprediksi 6 data.

3.2 Signifikansi Data

Pada penelitian yang pernah ada dinyatakan bahwa penggunaan data jenis kelamin sebagai variabel pelatihan dapat meningkatkan akurasi namun dalam pengujian pada alikasi ini penggunaan jenis kelamin sebagai variabel pelatihan tidak sesuai dengan penelitian yang pernah ada.

Tabel 2 Signifikansi data

Proporsi	Akurasi Metode tanpa Jenis kelamin		Akurasi Metode dengan jenis kelamin	
	C4.5	Naïve Bayes	C4.5	Naïve Bayes
50:50	0.84033	0.72268	0.79831	0. 70588
70:30	0.875	0.72222	0.90277	0. 77777
90:10	0.875	0.66666	0.875	0.70833

Berdasarkan peneltian yang telah dilakukan, perubahan nilai akurasi antara penggunaan variabel jenis kelamin tidak signifikan, dan tidak konstan. Ada dan tidak adanya varibel jenis kelamin dapat menaikkan nilai akurasi maupun menurunkan nilai akurasi. Beberapa hal tersebut dapat disebabkan oleh penyebaran data jenis kelamin yang tidak memadai, terjadi ketimpangan terhadap data jenis kelamin, tidak meratanya penyebaran jenis kelamin terhadap kelas kelulusan.

4. Kesimpulan

Kesimpulan yang didapatkan setelah melakukan pengujian terhadap aplikasi menggunakan metode C4.5 dan Naïve Bayes untuk prediksi masa studi mahasiswa adalah sebagai berikut:

1. Data pengujian yang tidak bersih dapat mempengaruhi hasil dimana outliers yang semestinya dihilangkan akan mengubah hasil prediksi. *Outliers* yang ikut dalam data

pelatihan akan mengubah model yang dihasilkan pada saat proses pelatihan, yang kemudian akan mempengaruhi proses pengujian.

2. Pada skenario pengujian yang dilakukan pembagian proporsi data pelatihan dan pengujian sebesar 50:50, 70:30, 90:10. Dari pengujian tersebut diperoleh nilai akurasi terbaik sebesar 0.90277 dengan proporsi data sebesar 70:30, akurasi terrendah sebesar 0.79831 dengan proporsi 50:50 dan 0.875 dengan proporsi sebesar 90:10. Dapat disimpulkan bahwa proporsi besarnya data training maupun data testing tidak menentukan nilai akurasi.
3. Variabel data pengujian dapat mempengaruhi model pelatihan. Penyebaran data variabel terhadap kelas hasil mempengaruhi model pelatihan.
4. Penambahan variabel jenis kemlamin tidak mengubah nilai akurasi secara signifikan.
5. Berdasarkan pengujian yang dilakukan menggunakan 237 data mahasiswa dapat dinyatakan bahwa metode C4.5 selalu unggul dibanding metode Naïve Bayes dalam setiap proporsi data dan pengujiannya.

REFERENSI

[1] Simon, S. and Trisnawarman, D., 2014. Aplikasi Prediksi Status Registrasi Mahasiswa Baru Menggunakan Metode Naïve Bayes dan Algoritma C4. 5. *Jurnal Ilmu Komputer dan Sistem Informasi*, *2*(2), pp.216-219.

[2] Kasmir dan Jakfar. Studi Kelayakan Bisnis, Edisi kedua,(Jakarta: Penerbit Kencana Prenada Media Group, 2007),h. 178

[3] Noir Primadona Purba , Masa Studi dan batas waktu, http://www.unpad.ac.id/pembelajaran/evaluasi-hasil-belajar-dan-batas-waktu-studi/masa-studi-dan-batas-waktu/, 6 Maret 2019

[4] .Administrasi akademik, Petunjuk Kegiatan Administasi Akademik & kegiatan Lainnya Tahun akademik 2018/2019, (Jakarta: UNTAR, 2016), hal. 12

[5] Solichin, Achmad. *Pemrograman web dengan PHP dan MySQL*. Penerbit Budi Luhur, 2016.

[6] Srinath, K. R. "Python–The Fastest Growing Programming Language." *International Research Journal of Engineering and Technology (IRJET) Volume* 4 (2017).

[7] Visa, Sofia, et al. "Confusion Matrix-based Feature Selection." *MAICS* 710 (2011): 120-127.

[8] Rohith Gandhi, *Naïve Bayes Classifier*, https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c, 6 Maret 2019

[9] Amin, Rafik Khairul, and Yuliant Sibaroni. "Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)." *2015 3rd International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2015.

Charles Yuliansen, Seorang mahasiswa pada program studi Fakultas Teknologi Informasi di Universitas Tarumanagara

Bagus Mulyawan, memperoleh gelar S.Kom. dari Universitas Gunadarma tahun 1992. Kemudian tahun 2008 memperoleh gelar M.M. dari Universitas Budi Luhur. Saat ini sebagai staf pengajar program studi Teknik Informatika Universitas Tarumanagara

Novario Jaya Perdana, memperoleh gelar S.Kom. dari Institut Teknologi Sepuluh November pada tahun 2011. Kemudian tahun 2016 memperoleh gelar M.T. dari Universitas Indonesia. Saat ini sebagai staf pengajar program studi Teknik Informatika, Universitas Tarumanagara