PERBANDINGAN ALGORITMA DATA MINING UNTUK PRESTASI BELAJAR MAHASISWA FIK

Green Ferry Mandias*[1], Green Arther Sandag[2], Kristop Nenoharan[3]
[1,2,3]Jurusan Informatika, FIK, UNKLAB, Manado
e-mail: *[1]green@unklab.ac.id, [2]greensandag@unklab.ac.id, [3]11210295@student.unklab.ac.id

Abstrak

Prestasi belajar merupakan salah satu aspek yang paling penting dalam bidang pendidikan dan menjadi harapan semua pihak. Bagi pihak perguruan tinggi prestasi belajar mahasiswanya merupakan salah satu indikator efektif proses belajar mengajar, yang sekaligus dapat digunakan untuk meningkatkan citra perguruan tinggi. Di perguruan tinggi prestasi belajar yang dicapainya oleh mahasiswa menggunakan Indeks Prestasi Kumulatif (IPK). Data demografi dan data akademik mahasiswa dapat digunakan dalam menganalisis dan memprediksi kinerja mahasiswa dengan menggunakan algoritma data mining diantaranya adalah Naive bayes, Decision Tree, K-Nearest Neighbor dan Random Forest. Dalam penelitian ini menggunakan 10-fold cross validation untuk memprediksi tingkat error dari data. Dataset dibagi dua yaitu data trainning dan data testing. Dari penelitian yang telah dilakukan maka didapatkan Random Forest memiliki tingkat akurasi yang paling tinggi dengan bobot 95.12%. Fitur yang paling berpengaruh adalah sekolah (SMA dan sederajat) dengan bobot 0.202. Berdasarkan hasil tersebut dapat disimpulkan bahwa ranndom forest memiliki accuracy yang paling tinggi dari ketiga algoritma lainnya.

Kata Kunci : *Naive bayes, Decision Tree, K-Nearest Neighbor, Random Forest.*

Abstract

Learning achievement is one of the most important aspects in the field of education and is the hope of all parties. For higher education parties, student learning achievement is one of the effective indicators of the teaching and learning process, which can also be used to improve the image of universities. In college the learning achievement achieved by students uses the Grade Point Average (GPA). Demographic data and student academic data can be used in analyzing and predicting student performance using data mining algorithms including Naive Bayes, Decision Tree, K-Nearest Neighbor and Random Forest. In this study using 10-fold cross validation to predict error rates from data. The dataset is divided into two, namely trainning data and testing data. From the research that has been done, the Random Forest has the highest accuracy with a weight of 95.12%. The most influential feature is the school (high school and equivalent) with a weight of 0.202. Based on these results it can be concluded that Ranndom Forest has the highest accuracy of the other three algorithms.

Keywords : *Naive bayes, Decision Tree, K-Nearest Neighbor, Random Forest.*

1. PENDAHULUAN

Fakultas Ilmu Komputer (FIK)Universitas Klabat berdiri sejak tahun 1999 pertama kali program studi ini bernama Ilmu Komputer, kemudian berubah menjadi Sistem Informasi. Pada tahun 2010 program studi ini sudah mendapatkan akreditasi dari Badan

Akreditasi Nasional (BAN) dengan nomor 021/BAN-PT/Ak-XIII/S1/X/2010 [1].

Fakultas Ilmu Komputer Universitas Klabat, memiliki dua program studi yaitu Sistem Informasi dan Informatika. Sistem informasi mendapat akreditisi B (339) dari BAN-PT (Badan Akreditasi Nasional perguruan tinggi) dengan nomor 972/SK/BAN-PT/Akred/S/IX/2015, Lulusan program studi Sistem Informasi akan menjadi seorang ahli dibidang Sistem Informasi, dimana mereka ini akan membuat aplikasi-aplikasi yang bermanfaat bagi manusia disegala bidang, seperti perkantoran, institusi, perhotelan, perdagangan dll. Sebuah sistem yang sebelumnya dikerjakan manual, oleh seorang ahli Sistem Informasi akan menjadi mudah, sederhana dan otomatis dengan bantuan komputer. Informatika mendapat akreditasi B (346) dari BAN-PT (Badan Akreditasi Nasional Perguruan Tinggi) nomor seri 149/SK/BAN-PT/Akred/S/VIII/2016, Seorang ahli informatika (Ilmu Komputer) harus mempunyai dasar algoritma dan logika yang kuat karena ia akan merancang suatu prosedur baru, modul baru ataupun algoritma baru [1].

Suatu universitas dikatakan berkualitas dapat dilihat dari beberapa faktor, satu diantarannya adalah prestasi mahasiswa, kemampuan mahasiswanya dilihat dari Indeks Prestasi Kumulatif (IPK) yang didapat selama masa kuliah sampai dengan lulus. Dari data akademik dapat memprediksi IPK mahasiswa dengan menggunakan teknik data mining.

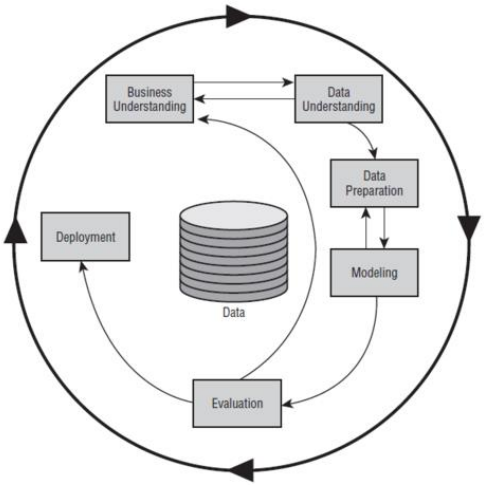
Prestasi belajar yang tinggi selalu menjadi harapan semua pihak. Bagi pihak universitas prestasi belajar mahasiswanya merupakan salah satu indikator efektif proses belajar mengajar, yang sekaligus dapat digunakan untuk meningkatkan citra suatu universitas tersebut. Data akademik, dan data geografis mahasiswa dapat digunakan dalam menganalisis dan memprediksi kinerja mahasiswa dengan menggunakan teknik-teknik data mining diantaranya adalah teknik decision tree, Naïve Bayes, random forest, dan knn. Dengan variabel-variabel penentu adalah umur saat masuk perguruan tinggi, jenis kelamin, suku bangsa, asal sekolah, sks yang diambil, pembiayaan kuliah, dan tempat tinggal [2].

Pada penelitian Ying Zhang dan Ving Zhang, menggunakan beberapa algoritma klasifikasi dalam data mining untuk mengetahui retensi (penyimpanan) mahasiswa menggunakan tiga algoritma dalam penelitiannya diantaranya Navie Bayes, Support Vectore Mechine dan Decision Tree untuk menganalisis prestasi akademik mahasiswa. Dalam penelitiannya didapat bahwa dari ketiga Algoritma Data Mining tersebut yang paling akurat adalah Navie Bayes dengan 89,5% dan yang kedua adalah Support Mechine dengan 83,5% dan yang terakhir adalah Decision Tree dengan 81,3% [3]. Dalam penelitian [4] mengatakan ketika dilakukan pengujian K Fold Cross validation sejumlah 10 Fold untuk dengan metode KNN pada data uji sejumlah 500 data, maka didapati nilai K terbaik adalah pada metode KNN adalah 5. Lebih lanjut [5] mengatakan sistem informasi data mining untuk siswa berpotensi akademik berbasis online dapat memberikan dukungan bagi orang tua dalam mengevaluasi potensi putra dan putrinya, dikarenakan kemudahan dalam melakukan akses informasi dan akurasi yang tinggi dari algoritma pohon keputusan.

Jadi masalah yang dapat dirumuskan yaitu bagaimana perbandingan Algoritma Data Mining terhadap prestasi belajar mahasiswa Fakultas Ilmu Komputer UNKLAB berdasarkan variabel demografi yaitu: (suku, pembiayaan, bekerja, tempat tinggal, umur, jenis kelamin) dan akademik yaitu: (sks, IPK (indeks prestasi kumulatif)). Tujuan dari penelitian ini yaitu mengetahui perbandingan Algoritma Data Mining terhadap prestasi belajar mahasiswa Ilmu Komputer UNKLAB.

2. METODE PENELITIAN

Pada penilitian ini penulis memilih proses model CRISP-DM (Cross Industry standard processfor data mining) [6] seperti pada gambar 1.



Gambar 1 Metode Penelitian

Model ini menjadi dasar teori dalam metode penelitian ini. Adapun langkah-langkah dalam penelitian ini yaitu :

1. Fase Business Understanding, data yang diperoleh dari Mahasiswa FIK UNKLAB selama ini belum pernah dilakukan pengambilan data tentang pembiayaan kuliah terhadap data tersebut dan belum dimanfaatkan dalam menentukan prediksi prestasi akademik mahasiswa menggunakan metode Data Mining. Oleh karena itu penelitian ini akan menggali data tersebut dengan menggunakan Algoritma Data Mining.
2. Fase Data Understanding, data yang akan diambil dari mahasiswa FIK UNKLAB adalah suku, sekolah, bekerja, tempat tinggal, umur, jenis kelamim, sks yang diambil tiap semester, ipk (indeks prestasi kumulatif.)
3. Fase Data Preparation, Dari data mahasiswa FIK UNKLAB maka akan dilakukan teknik data preparation agar kualitas data diperoleh lebih baik dengan cara :

    a) Data Validation, untuk mengidentifikasi dan menghapus data yang ganjil, data yang tidak konsisten dan data yang tidak lengkap.
    b) Data Integration and Transformation : untuk meningkatkan akurasi dan efisensi algoritma. Data yang digunakan dalam penelitian ini bernilai kategorikal untuk model klasifikasi, data ditransformasi ke dalam angka menggunakan software Weka.
    c) Data Size Reduction dan Dicretization : untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informative. Di dalam data training yang digunakan dalam penelitian ini,dilakukan seleksi atribut dan penghapusan data duplikasi.
4. Fase Modeling, pada tahap ini data di proses dan diklasifikasikan sehingga menghasilkan sejumlah aturan dengan mengunakan algoritma data mining seperti penelitian [7]
5. Evaluation Phase, pada fase ini dilakukan pengujian data terhadap model klasifikasi dengan algoritma data mining. [8]

2.1 Instrumentasi Penelitian

Dalam Penelitian ini penulis menggunakan sumber data sekunder data sekunder merupakan data yang diperolehkan dengan cara mengambil data mahasiswa dari sistem informasi UNKLAB, melakukan tinjauan pustaka, membaca buku-buku, artikel maupun jurnal dengan mengambil referensi dari sumber-sumber pustaka yang berkaitan dengan penelitian ini. Studi pustaka juga dilakukan dengan menggunakan internet dan observasi.

2.2 Lokasi Penelitian

Penelitian dilaksanakan di Fakultas Ilmu Komputer Universitas Klabat berlokasi di jln. Arnold Mononutu, Airmadidi, Minahasa Utara, Sulawesi Utara

2.3 Peralatan

Peralatan dalam penelitian ini meliputi kebutuhan perangkat lunak dan kebutuhan perangkat keras. Dibawah ini merupakan kebutuhan dari sistem, diantaranya:

Kebutuhan perangkat lunak :

1. Microsoft Office Word 2013, software ini digunakan untuk mengolah laporan hasil penelitian
2. Sistem operasi Windows 7, sistem Operasi yang digunakan dalam notebook penulis.
3. Rapidminer, software yang digunakan untuk melihat hasil akurasi dari algoritma yang digunakan terhadap datasheet yang diteliti.

Kebutuhan Perangkat Keras :

1. Processor : Intel(R) Core(TM) i3-2310M CPU @ 2.10GH2 2.10 GH2
2. Ram 2 gb
3. Harddisk : 500 gb
4. Satu buah mouse
5. Satu buah printer canon

3. HASIL DAN PEMBAHASAN

Populasi mahasiswa Fakulatas Ilmu Komputer Universitas Klabat berjumlah 540 mahasiswa, tingkat I berjumlah 140 mahasiswa dan tingkat II, III, IV berjumlah 300 mahasiswa. Dalam penelitian ini hanya memgambil data tingkat II, III dan IV. Karena data yang diambil semester II tahun ajaran 2017/2018. Sampel yang digunakan dalam penelitian ini berjumlah 205 mahasiswa dan data yang diambil adalah suku, SMA dan sederajat, umur saat masuk kuliah, sks, ipk, tempat tinggal, pembiyaan kuliah, penghasilan orangtua, penghasilan mahasiswa, bekerja atau tidak.

Data menunjukkan atribut suku bangsa yang paling banyak mahasiswa berasal dari Minahasa dengan jumlah 123 mahasiswa, atribut sekolah (SMA dan sederajat) yang paling banyak mahasiswa tamatan SMA dengan jumlah 166 mahasiswa, pembiayaan selama kuliah yang paling banyak membiayayai adalah orangtua dengan jumlah 187 mahasiswa, bekerja (labor/tidak) yang paling banyak adalah mahasiswa tidak bekerja dengan jumlah 162 mahasiwa, penghasilan orangtua selama 1 bulan yang paling banyak penghasilan orangtua adalah 2-4 juta dengan junlah 92 mahasiswa, penghasilan mahasiswa selama 1 bulan yang paling banyak penghasilan mahasiwa adalah <2 juta dengan jumlah 163 mahasiswa, tempat tinggal selama kuliah yang paling banyak tinggal dikos dengan jumlah 127 mahasiswa, umur saat masuk kuliah yang paling banyak adalah umur <18 tahun dengan jumlah 152 mahasiswa, sks yang diambil tiap semester yang paling banyak mahasiswa yang mengambil sks 18-21 sks dengan jumlah 130 mahasiswa, indeks prestasi kumulatif (IPK) yang paling banyak IPK >3,00 dengan jumlah 139 mahasiswa.

3. 1 Algoritma Naive Bayes

Tabel 1 Cross Validation Naive Bayes

	true High	true Normal	true Low	class precision
pred.High	93	32	3	72.66%
pred.Normal	17	14	4	40.00%
pred.Low	1	0	0	0.00%
class recall	83.78%	30.43%	0.00%	

Pada tabel 1 Cross validation Naive Bayes. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 72.66%, pred.Normal 40.00% dan pred.Low 0.00%. Class recall : true High 83.78%, true Normal 30.43% dan true Low 0.00%.

Tabel 2 Independent Naive Bayes

	true High	true Normal	true Low	class precision
pred.High	88	27	6	72.73%
pred.Normal	20	16	1	43.24%
pred.Low	3	3	0	0.00%
class recall	79.28%	34.78%	0.00%	

Pada tabel 2 Independent Naive Bayes. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 72.73%, pred.Normal 43.24% dan pred.Low 0.00%. Class recall : true High 79.28%, true Normal 34.78% dan true Low 0.00%.

3.2 Algoritma Decision Tree

Tabel 3 Cross Validation Decision Tree

	true High	true Normal	true Low	class precision
pred.High	91	34	4	70.54%
pred.Normal	20	10	3	30.30%
pred.Low	0	2	0	0.00%
class recall	80.98%	21.74%	0.00%	

Pada tabel 3 Cross validation Decision Tree. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 70.54 pred.Normal 30.30% dan pred.Low 0.00%. Class recall : true High 80.98%, true Normal 21.74% dan true Low 0.00%.

Tabel 4 Independent Decision Tree

	true High	true Normal	true Low	class precision
pred.High	27	2	0	93.10%
pred.Normal	1	9	0	90.00%
pred.Low	0	0	2	100.00%
class recall	96.43%	81.82%	100.00%	

Pada tabel 4 Independent Decision Tree. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 93.10%, pred.Normal 90.00% dan pred.Low 100.00%. Class recall : true High 96.43%, true Normal 81.82% dan true Low 1000.00%.

3.3 Algorimta K-Nearest Neighbor

Tabel 5 Cross Validation K-Nearest Neighbor

	true High	true Normal	true Low	class precision
pred.High	91	34	4	70.54%
pred.Normal	20	10	3	30.30%
pred.Low	0	2	0	0.00%
class recall	80.98%	21.74%	0.00%	

Pada tabel 5 Cross Validation K-Nearest Neighbor. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 70.54%, pred.Normal 30.30% dan pred.Low 0.00%. Class recall : true High 80.98%, true Normal 21.74% dan true Low 0.00%.

Tabel 6 Independent K-Nearest Neighbor

	true High	true Normal	true Low	class precision
pred.High	26	3	1	86.67%
pred.Normal	2	8	1	72.73%
pred.Low	0	0	0	0.00%
class recall	92.86%	72.73%	0.00%	

Pada tabel 6 Independent K-Nearest Neighbor. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 86.67%, pred.Normal 72.73% dan pred.Low 0.00%. Class recall : true High 92.86%, true Normal 72.73% dan true Low 0.00%.

3.4 Algorimta Random Forest

Tabel 7 Cross Validation Random Forest

	true High	true Normal	true Low	class precision
pred.High	89	32	4	71.20%
pred.Normal	21	13	3	35.14%
pred.Low	1	1	0	0.00%
class recall	80.18%	28.26%	0.00%	

Pada tabel 7 Cross validation Random Forest. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 71.20%, pred.Normal 35.14% dan pred.Low 0.00%. Class recall : true High 80.18%, true Normal 28.26% dan true Low 0.00%.

Tabel 8 Independent Random Forest

	true High	true Normal	true Low	class precision
pred.High	27	1	0	96.43%
pred.Normal	1	10	0	90.91%
pred.Low	0	0	2	100.00%
class recall	96.43%	90.91%	100.00%	

Pada tabel 8 Independent Random Forest. Dalam perhitungan tabel menggunakan confusion matrix, Class precision : pred.High 96.43% pred.Normal 90.91% dan pred.Low 100.00%. Class recall : true High 96.43%, true Normal 90.91% dan true Low 100.00%.

3.5 Perbandingan Cross Validation

Tabel 9 Cross Validation

Cross Validadation	Accuracy	Recall	Presisi
Naive Bayes	65.24%	38.07%	37.55%
Decision Tree	61.59%	34.57	33.62%
k-NN	62.20%	33.60%	32.28%
Random Forest	62.20%	36.15%	35.45%

Pada tabel 9 Cross Validattion Accuracy yang paling tinggi ialah Naive Bayes dengan bobot 65.24%, karena naive bayes memiliki data yang lebih banyak pada tabel 4.3 pred.High (93), pred.Normal (17) dan pred.low (1). Recall yang paling tinggi ialah Random Forest dengan bobot 95.78%. Dan Presisi yang paling tinggi ialah Random Forest dengan bobot 95.78%. Decision tree memiliki akurasi yang lebih rendah 61.59%, karena data yang lebih sekikit, pada tabel 4.5 pred.high (91), pred.normal (20), dan pred.low (0). Random forest memiliki acuracy 62.20% karena pada tabel 4.4 pred.high (89), pred.normal (21), dan pred.low (1). K-nearest neigbor memiliki acuracy 62.20% , karena pada tabel 4.7 pred.high (91), pred.normal (20), dan pred.low (0)

3.6 Perbandingan Independent

Tabel 10 Independent

Independent	Accuracy	Recall	Presisi
Naive Bayes	63.41%	38.02%	38.66%
Decision Tree	92.68%	92.75%	94.37%
k-NN	82.93%	55.19%	53.13%
Random Forest	95.12%	95.78%	95.78%

Pada tabel 10 Independent Accuracy yang paling tinggi ialah Random Forest dengan bobot 95,12%, karena memiliki data yang lebih sedikit, karena pada tabel 4.10 pred.high (27), pred.normal (1), dan (0).Recall yang paling tinggi ialah Random Forest dengan bobot 95.78%. Dan Presisi yang paling tinggi ialah Random Forest dengan bobot 95.78%. Naive bayes memiliki acuracy yang lebih rendah 63.41% dari ketiga algortima lainya, karena memiliki data yang lebih banyak kerena pada tabel 4.4 pred.high (88), pred.normal (20) dan pred.low (3). Decision tree memiliki acuracy 92.68%, karena pada tabel 4.6 pred.high (27), pred.normal (1) dan pred low (0). K-nearest neigbor memiliki acuracy 82.93% karena pada tabel 4.8 pred.high (26), pred.normal (2), dan pred.low (0)

4. KESIMPULAN

Kesimpulan yang dapat diambil dari penelitian tentang perbandingan algoritma data mining terhadap prestasi belajar mahasiswa FIK UNKLAB ditinjau dari aspek akademik dan ekonomi adalah sebagai berikut :

1. Penelitian ini menggunakan 10-fold cross validation untuk memprediksi tingkat error dari data. Dataset dibagi dua yaitu data trainning dan data testing. Cross validation dengan accuracy yang paling tinggi adalah naive bayes 65.24%, k-nn 62.20%, random forest 62.20%, decision tree 61.59%.
2. Independent dengan accuracy yang paling tinggi adalah random forest 95.12%, decision tree 92.68%, k-nn 82.93%, naive bayes 63.41%
3. Dalam penelitian ini ada sepuluh atribut yang digunakan dalam penelitian ini, ada tiga atribut dengan bobot yang paling tinggi adalah sekolah (SMA dan sederajat) : 0.209, tempat tinggal selama kuliah : 0.200, jenis kelamin : 0.199.
4. Dalam penelitian ini menggunakan empat algoritma : naive bayes desision tree, random forest, k-nearest neigbor. Algoritma dengan bobot yang paling tinggi adalah random forest dengan accuracy 95.12%.

5. SARAN

Untuk keperluan penelitian lebih lanjut mengenai perbandingan metode klasifikasi data mining, dapat dilakukan pengembangan untuk dapat menghasilkan model yang lebih baik lagi serta menggunakan algoritma pengklasifikasi lain yang mungkin diluar supervised learning agar dapat dilakukan penelitian yang berbeda dari umumnya yang sudah ada.

DAFTAR PUSTAKA

[1] http://www.unklab.ac.id/id/fakultas-ilmu-komputer.

[2] S. Defiyanti, "Perbandingan: Prediksi Prestasi Belajar Mahasiswa Menggunakan Teknik Data Mining (Study Kasus Fasilkom Unsika)," Konferensi Nasional Sistem Informasi 2014,, p. 117, 2014.

[3] Y. Zhang, S. Oussena, T. Clark and H. Kim, "Use Data Mining To Improve Student Retention In Higher Education – A Case Study," 2014.

[4] M. Khlil., Kusrini. , Henderei. Penerapan Metode K Nearest Neighbord Dalam Proses Seleksi Penerima Beasiswa, “Seminar Nasional Sistem Informasi dan Teknologi Informasi”, p. 13-18, 2018

[5] D. Setiyadi., A. Nurdin. ”Data Mining Potensi Akademik Siswa Berbasis Online”. Jurnal Sisfotenika. ISSN: 2460-5344, Vol. 2, No. 1, 2012

[6] D. Feblian and D. U. Daihani, "Implementasi Model Crisp-Dm Untuk Menentukan Sales Pipeline Pada Pt X," Jurnal Teknik Industri ISSN: 1411-6340, p. 5, 2015.

[7] D. Sofi , "Perbandingan: Prediksi Prestasi Belajar Mahasiswa Menggunakan Teknik Data Mining (Study Kasus Fasilkom Unsika)," Konferensi Nasional Sistem Informasi 2014,, p. 118, 2014.

[8] A. R. Khadafy, "Penerapan Naive Bayes untuk Mengurangi Data Noise pada Klasifikasi Multi Kelas dengan Decision Tree," Konferensi Nasional Sistem Informasi 2014, vol. 2, p. 136, 2015.