Pembentukan Model Keputusan Untuk Penentuan Pola Pengetahuan Seleksi Calon Mahasiswa Jalur PMDK Dengan Artificial Neural Network

Tacbir Hendro Pudjiantoro [1)], Erna Piantari [2)] Asri Maspupah [3)]
Jurusan Informatika, Fakultas Sains dan Informatika
Universitas Jenderal Achmad Yani
Jl. Terusan Sudirman, Cimahi
tacbir23501027@yahoo.com, erna.piantari14@gmail.com asri.maspupah89@gmail.com

Abstrak

Jalur PMDK (Penelusuran Minat dan Kemampuan) adalah salah satu sistem penerimaan mahasiswa baru yang digunakan oleh perguruan tinggi. Jalur ini mempunyai kriteria penerimaan mahasiswa berdasarkan nilai raport atau prestasi akademik yang dimiliki oleh kandidat mahasiswa. Kenyataannya, selain nilai raport atau prestasi akademik, ada hal lain yang harus dipertimbangkan untuk menentukan peluang keberhasilan sebagai mahasiswa yang lulus tepat waktu. Jumlah atribut data yang banyak dan data yang tidak seimbang menjadi permasalahan untuk dapat menentukan atribut mana saja yang menggambarkan permasalahan yang sebenarnya. Penelitian ini membentuk pola yang digunakan menggunakan PCA (Principal Component Analysis) dan SMOTE untuk menentukan atribut mana saja yang dapat digunakan sebagai penentu sehingga dapat menggambarkan kondisi mahasiswa yang sebenarnya, untuk selanjutnya berdasarkan atribut tersebut dapat dibobotkan dan kemudian dihitung hasil pembobotannya. Nilai pembobotan tersebut kemudian dibandingkan dengan nilai minimal yang dikendaki, jika lebih besar maka kandidat mahasiswa tersebut dapat diterima. Penelitian ini berhasil membentuk pola yang dapat digunakan untuk memprediksi kandidat mahasiswa melalui jalur PMDK, namun demikian masih perlu diuji keakuratannya.

Kata kunci: *Principal Component Analysis, SMOTE, Data Mining*

1. Pendahuluan

Jalur PMDK (Penelusuran Minat dan Kemampuan) adalah sistem penerimaan mahasiswa baru yang diselenggarakan oleh perguruan tinggi secara mandiri. Jalur ini menggunakan kriteria penerimaan mahasiswa berdasarkan nilai rapot atau prestasi akademik yang dimiliki oleh kandidat mahasiswa. Namun pada kenyataannya selain nilai rapot, banyak hal yang harus dipertimbangkan untuk menentukan apakah kandidat mahasiwa tersebut memiliki peluang keberhasilan sebagai mahasiswa yang lulus tepat waktu. Permasalahan lain yaitu sebagian besar kandidat mahasiswa yang menggunakan jalur PMDK sebagai sebuah batu loncatan saja agar diterima sebagai mahasiswa sehingga memiliki peluang keberhasilan yang kecil untuk mengikuti proses pembelajaran sampai akhir. Ketersediaan data history kelulusan mahasiswa jalur PMDK menjadi salah satu modal utama untuk mengetahui kriteria mahasiswa yang memiliki kemungkinan besar sukses dalam pendidikannya dengan menggunakan jalur tersebut.

Permasalahan yang ada adalah jumlah data yang banyak dari peserta jalur PMDK, sementara kandidiat mahasiswa yang diterima sebanyak 10% dari daya tampung yang disediakan. Masalah selanjutnya adala Jumlah Atribut yang banyak untuk setiap kandidiat mahasiswa, dengan kondisi tidak semua atribut tersebut terisi dengan benar.

Manfaat dari penelitian ini berdasarkan permasalahan di atas adalah untuk meminimalkan atribut yang besar dengan melakukan ektraksi fitur, dan memaksimalkan nilai dari atribut yang digunakan. Pola yang terbentuk dapat digunakan untuk melakukan prediksi kandidat mahasiswa jalur PMDK yang diperkirakan akan mengikuti kuliah sampai dengan wisuda. Apabila pola prediksi yang terbentuk mempunyai kondisi ideal, maka akan dapat mengurangi tingkat "*Drop Out*" mahasiswa.

2. Metode Penelitian

PCA (Principal Component Analysis)

Principal component analysis merupakan suatu teknik matematika yang biasa digunakan untuk mengidentifikasikan pola dari sebuah data set dan mengekspesikan data sedemikain rupa untuk mendapatkan kesamaan dan perbedaan yang terdapat pada dataset tersebut [1]. Selain itu, teknik ini dapat bekerja dengan mengubah sejumlah variable yang memiliki korelasi menjadi sejumlah variable kecil yang disebut sebagai komponen utama[2].

PCA menggunakan sebuah vector space transforms untuk mengurangi dimensi suatu data set, dengan tujuan PCA adalah [3] :

1. Mengekstraksi informasi penting sari sebuah dataset
2. Melakukan kompresi dataset dengan ukuran menjadi lebih keci dan hanya mempertahankan informasi penting
3. Merubah bentuk data menjadi lebih sederhana
4. Melakukan analisis struktur variable data

SMOTE

Confusion matrix

Confusion Matrix adalah metode yang digunakan untuk melakukan evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas yang asli dari inputan atau dengan kata lain berisi informasi nilai aktual dan prediksi pada klasifikasi.[4]

Tabel 1 . *Confusion Matrix*

Classification	Predicted Class	
	Class = Yes	Class=No
Class = Yes	*true positive*-TP	*false negative*-FN
Class = No	*false positive*-FP	*true negative*-TN

Rumus untuk menghitung tingkat akurasi berdasarkan nilai matriks tersebut adalah :

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FN}$$

Artificial Neural Network

Artificial Neural Network (ANN) atau disebut juga sebagai jaringan saraf tiruan. Salah satu metoda ANN yang akan diterapkan dalam penelitian ini adalah jaringan Hebb.

Jaringan Hebb yaitu model *neuron* dengan meghitung bobot dan bias secara iteratif. Dasar dari algoritma Hebb adalah kenyataan bahwa apabila 2 *neuron* yang dihubungkan dengan sinapsis secara serentak menjadi aktif (sama-sama bernilai positif atau negatif), maka kekuatan sinapsisnya meningkat. Sebaliknya, apabila kedua *neuron* aktif secara tidak sinkron (salah satu bernilai positif dan yang lainnya bernilai negatif), maka kekuatan sinapsisnya melemah.[6]

Karena itulah, dalam setiap iterasi, bobot sinapsis dan bias diubah berdasarkan perkalian *neuron-neuron* dikedua sisinya. Untuk jaringan layar tunggal dengan 1 unit keluaran dimana semua unit masukan x_i terhubung langsung dengan unit keluaran y, maka perubahan nilai bobot dilakukan berdasarkan persamaan :

$w_i(baru) = w_i(lama) + x_i * y$

dengan:

w_i : bobot data input ke-i;

x_i : input data ke-i.

Algoritma pelatihan Hebb dengan vektor input s dan unit target t adalah sebagai berikut :

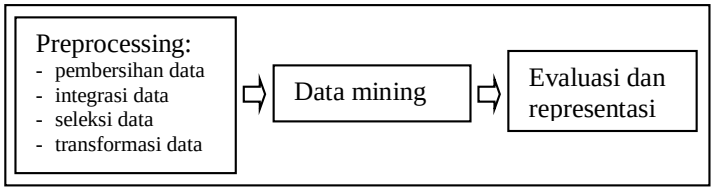
1. Inisialisasi semua bobot = $w_i = 0$ (i = 1, ... , n)
2. Untuk semua vektor input s dan unit target t, lakukan :
   a. Set aktivasi unit masukan $x_i = s_i$ (i = 1, .., n)
   b. Set aktivasi unit keluaran : y=t
   c. Perbaiki bobot menurut persamaan :
      $W_i(baru) = w_i(lama) +$ w (i=1, ..., n)
   d. Perbaiki bias menurut persamaan b(baru) = b(lama)

Perhatikan bahwa perbaikan bias diperlakukan sama seperti bobot.

Masalah yang sering timbul dalam jaringan Hebb adalah dalam menentukan representasi data masukan/ keluaran untuk fungsi aktivasi yang berupa treshhold. Representasi yang sering dipakai adalah bipolar (nilai -1 atau 1). Kadangkala jaringan dapat menentukan pola secara benar jika dipakai representasi bipola saja, dan akan salah jika dipakai representasi biner (nilai 0 atau 1).

3. Hasil dan Pembahasan

Pembangunan model keputusan yang dilakukan dalam penelitian ini melibatkan enam proses sekuensial data mining [5][6] yaitu proses pembersihan data, proses integrasi data, proses seleksi data, proses transformasi data, proses data mining serta proses evalusi dan representasi hasil. Empat proses pertama dalam proses sekuansial tersebut termasuk kedalam tahapan *preprocessing* yang dapat dilihat pada Gambar 1.



Gambar 1. Proses Sequensial Data Mining

Preprocessing

Tahapan *preprocessing* dilakukan untuk menyiapkan data yang akan diproses untuk pembangunan model, tahapan ini terdiri dari empat proses.

Data yang digunakan terdiri dari dua sumber data, yaitu data rekap pendaftaran dan data kelulusan. Data rekap pendaftaran merupakan data identitas calon mahasiswa, data nilai rapot SMA dan data hasil Ujian Akhir Nasional tingkat SMA. Data tersebut diperoleh pada saat calon mahasiswa mendaftar. Data kelulusan merupakan rekap data yang berisi nilai yang diperoleh mahasiswa selama menempuh pendidikan di universitas sampai dengan selesai atau berhenti kuliah.

Proses pembersihan data dilakukan untuk membuang *noise* dan data yang tidak konsisten. Tahap ini dilakukan untuk memastikan bahwa data yang akan diproses merupakan data yang benar.

Ekstraksi fitur dengan PCA

Ekstraksi fitur merupakan tahapan dalam *preprocessing* yang digunakan untuk mengurangi jumlah fitur data tanpa kehilangan infromasi dari data tersebut. Dalam penelitian ini metode PCA (*Principle Component Analysis*) digunakan untuk melakukan proses ekstrasi fitur data. Data awal memiliki 26 fitur. PCA telah berhasil melakukan proses ekstraksi menjadi 15 fitur baru dengan nilai standar deviasi terbesar adalah 3.048 dan nilai standar deviasi terkecil adalah 0.519.

Dari 15 fitur hasil ekstraksi, dipilih hanya fitur yang memiliki standar deviasi memenuhi batas *threshold*. *Threshold* yang dipilih adalah setengah dari nilai standar deviasi terbesar, yaitu 1.52, sehingga fitur yang terpilih berjumlah 11 fitur.

Tabel 2. Hasil Proses ekstraksi dengan PCA

Feature	Stdev	Feature	Stdev
1	3.048	9	1.683
2	2.372	10	1.616
3	2.258	11	1.569
4	2.137	12	1.437
5	1.981	13	1.381
6	1.901	14	1.359
7	1.806	15	1.239
8	1.785		

SMOTE

Selain jumlah fitur data yang banyak, jumlah *instance data* yang terbatas menjadi permasalahan dalam proses pembangunan model pembelajaran. *SMOTE* adalah metode yang dapat digunakan untuk menciptakan data.[7]

Data Mining

Proses *data mining* merupakan proses utama yang dilakukan untuk menghasilkan *knowledge* dari

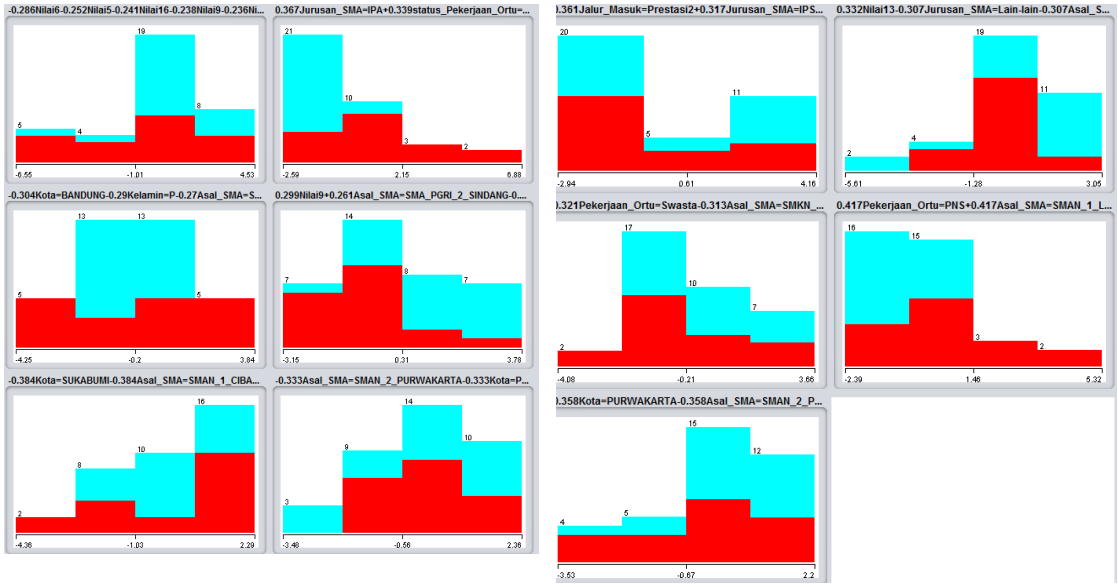
data yang dikumpulkan. *Knowledge* yang dihasilkan dari proses ini dapat berupa sebuah pola atau sebuah persamaan yang dapat dijadikan sebagai sebuah model untuk melakukan prediksi. Dalam penelitian ini metode klasifikasi dengan Evaluasi dan Refresentasi Model

Proses evaluasi dilakukan untuk mendapatkan nilai *performa* dari sebuah model. Teknik yang digunakan dalam evaluasi adalah dengan menggunakan metode *confusion matrix*.

Implementasi dan pembahasan hasil diskusi dengan tools Weka 3.8.1.

SMOTE (Resample Dataset by Applying the Systhetic Minority Oversampling Technique)

Hasil resample dataset dengan menggunakan SMOTE adalah eperti pada gambar 2.



Gambar 2. Resample Dataset

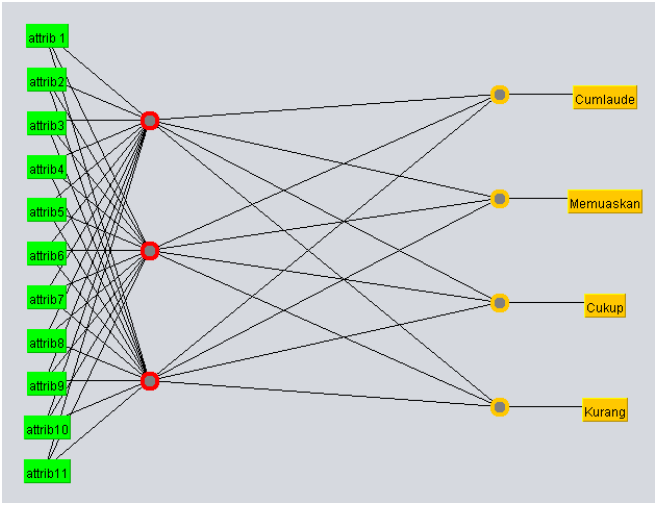
Training ANN

Hasil Training dengan menggunakan data hasil ekstraksi hasil menjadi seperti pada tabel 3.

Tabel 3. Arsitektur Model Jaringan Neural Network

Arsitektur	Jumlah Epoch	Rate Learning	Momentum	Correctly Classified Instance	TP Rate (True Positive)
11-1-4	500	0.3	0.2	75%	0.75
11-2-4	500	0.3	0.2	83.33%	0.833
11-3-4	500	0.3	0.2	86.11%	0.861
11-4-4	500	0.3	0.2	80.56%	0.806
11-5-4	500	0.3	0.2	80.56%	0.806
11-6-4	500	0.3	0.2	80.56%	0.806
11-7-4	500	0.3	0.2	80.56%	0.806
11-8-4	500	0.3	0.2	80.56%	0.806
11-9-4	500	0.3	0.2	80.56%	0.806
11-4-8-4	500	0.3	0.2	80.56%	0.806
11-4-4-4	500	0.3	0.2	77.78%	0.778
11-8-8	500	0.3	0.2	69.44%	0.698

Dari tabel hasil training tersebut diperoleh arsitektur Jaringan Neural Network terbaik dengan nilai Correctly Calssified Instance 83.33% dan Nilai True Positive Rate 0.833 dengan arsitektur 11 node input, 1 hiden layer yang terdiri dari 3 node dan 4 node output. Gambar 3.



Gambar 3. Arsitektur Jaringan Neural Network terbaik

Dengan input dan nilai bobot setiap node seperti pada tabel 4.

Tabel 4. input dan nilai bobot setiap node

Node	Input		Node	Input	
Node 0 (Class Cumlaude)	Threshold	-2.8040930565075746	Node 5	Threshold	1.359448733801175
	Node 4	-1.586335863368172		Attrib1	0.9587301968893015
	Node 5	-1.3517247656826417		Attrib2	-1.8886578493967021
	Node 6	-1.4134205866727696		Attrib3	1.7981893243643856
Node 1 (Class Memuaskan)	Threshold	4.223581111478451		Attrib4	0.028781573981739222
	Node 4	5.232472153489514		Attrib5	-4.362416448045569
	Node 5	-5.20041025212682		Attrib6	-0.1562073959282111
	Node 6	-2.7157682576453452		Attrib7	-2.4109212398234097
Node 2 (Class Cukup)	Threshold	-4.213095970439566		Attrib8	-2.6298215015248068
	Node 4	-5.239453299461243		Attrib9	-1.8739814052577977
	Node 5	5.196954084293966		Attrib10	0.4765706464181696
	Node 6	2.707257346606386		Attrib11	1.7210282547523192
Node 3 (Class Kurang)	Threshold	-2.7810756067262514	Node 6	Threshold	-1.3180262623143115
	Node 4	-1.6033211662934639		Attrib1	0.7933756380981883
	Node 5	-1.4179954365493839		Attrib2	-3.5191632923603615
	Node 6	-1.3804433282974806		Attrib3	-0.08204265629850811
Node 4	Threshold	1.193201741294053		Attrib4	0.6561899832348184
	Attrib1	-0.9038314442100785		Attrib5	-0.8747136383246894
	Attrib2	3.33360145045178		Attrib6	5.454443134557757
	Attrib3	-0.2600926760659297		Attrib7	1.440437766183051
	Attrib4	-0.5368156564804721		Attrib8	-2.272965992040895
	Attrib5	1.1847509302098862		Attrib9	-0.12528576480905496
	Attrib6	-5.109437248047133		Attrib10	0.47906114094680474
	Attrib7	-1.0736943956090212		Attrib11	1.3710485214007853
	Attrib8	2.519637526625979			
	Attrib9	0.7373523627574287			
	Attrib10	-0.6560907690022982			
	Attrib11	-1.3050344771249234			

Evaluasi Confusion Matrix

```
 a  b  c  d   <-- classified as
 0  0  0  0 |  a = Cumlaude
 0 15  3  0 |  b = Memuaskan
 0  3 15  0 |  c = Cukup
 0  0  0  0 |  d = Kurang
```

4. Simpulan

Pada penelitian ini, telah dibangun sebuah model pembelajaran untuk membangun pola pengetahuan yang melibatkan *dataset* dengan jumlah atribut yang banyak dan jumlah *instance data* yang terbatas. Untuk menangani pembangunan model pembelajaran yang menggunakan jumlah *dataset* yang kecil dapat dilakukan dengan mengurangi jumlah atribut data dari data set tersebut tanpa kehilangan informasi yang terdapat dalam data tersebut. Metode PCA dapat digunakan untuk melakukan proses ekstraksi atribut data sehingga dapat mengurangi jumlah atribut data tanpa kehilangan informasi dari data yang digunakan. Metode ini telah berhasil mengurangi jumlah atribut data dari 26 atribut menjadi 11 atribut. Selain itu, teknik SMOTE digunakan untuk menciptakan sintentik yang digunakan sebagai data *instance dataset* baru. Dengan menggunakan hasil dari preprocessing data dengan PCA-SMOTE proses pembangunan model pembelajaran digunakan dengan Algoritma Artificial Neural Network (ANN). Dari hasil pembelajaran tersebut dihasilkan model ANN terbaik yaitu ANN dengan arsitektur 11 node input, 1 hiden layer yang terdiri dari 3 node dan 4 node output. Evaluasi terhadap model yang dihasilkan dilakukan dengan menggunakan Confusion Matrix dengan nilai rate true positive sebesar 0.833 dan Correctly Calssified Instance 83.33%. Sedangkan model pembelajaran tanpa PCA-SMOTE menghasilkan nilai rate true positive sebesar 0.611 dan Correclty Classified Instance sebesar 61.1%. Dari penelitian yang tersebut maka dapat dikesimpulkan bahwa penggunaan PCA-SMOTE dapat digunakan untuk memperbaiki proses pembelajaran yang melibatkan dataset dengan atribut yang banyak dan jumlah instance data yang sedikit.

Daftar Pustaka

[1] Smith Lindsay, "A tutorial on Principal Components Analysis" , Feb 26, 2002, http://www.cs. otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. Springer.

[2] Abdi Herve, Lynne J. Williams. "Principal Compenent Analysis". Volume 2, July/August 2010, John Wiley & Sons, Inc.

[3] Richardson mark, "Principal Component Analysis", May 2009. URL:http://people.maths.ox.ac. uk/ richardsonm/SignalProcPCA.pdf

[4] Landgrebe Thomas C.W., Duin Robert P.W., "Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis" Volume: 30, May 2008, IEEE Transactions on Pattern Analysis and Machine Intelligence

[5] Nisbet, Robert; Elder, John; Miner, Gary; *Handbook of Statistical Analysis & Data Mining Applications* (2009);, Academic Press/Elsevier

[6] Chawla, Nitesh V. *Data Mining and Knowledge Discovery Handbook*, Springer; 2010 In: Maimon, Oded; Rokach, Lior (Eds)

[7] Davis,D.N. Rahman,M.M. "Addressing the Class Imbalance Problem in Medical Datasets" vol. 3, no. 2, 2013, International Journal of Machine Learning and Computing.

[8] Naseriparsa, Mehdi & Mansour Riahi Kashani, Mohammad. "Combination of PCA with SMOTE Resampling to Boost the Prediction Rate in Lung Cancer Dataset"; Maret 2014; International Journal of Computer Applications.

[9] Noori Roohollah , reza Abdul, Karbassi, Sabahi Mohammad Salman , "Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction", Volume 91, Issue 3, January–February 2010, Journal of Environmental Management.

[10] Noori Roohollah, Khakpour Amir, Omidvar Babak, Farokhnia Ashkan; "Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic", Volume 37, Issue 8, August 2010, Pages 5856-5862, Journal Expert Systems with Applications