POLA KEMAMPUAN ANAK BERDASARKAN RAPOR MENGGUNAKAN TEXT MINING DAN KLASIFIKASI NEAREST NEIGHBOR

Putri Eka Prakasawati[1)], Gunawan Abdillah[2)], Asep Id Hadianan[3)]

[1), 2)] Program Studi Informatika MIPA Unjani Cimahi
[3)] Program Studi Informatika MIPA Unjani Cimahi
Jln. Trsn. Jenderal Sudirman, Cimahi 40513. Gd. Lab II F-MIPA. PO BOX 148 Cimahi, Jawa Barat
Email : ekaprakasaputri@gmail.com[1)], abizakkiyy@y.ac.id[2)], ahadiana@gmail.com[3)]

Abstrak

Kemampuan merupakan kecakapan atau potensi seseorang untuk menguasai keahlian dalam melakukan sebuah pekerjaan yang beragam ataupun suatu penilaian atas tindakan seseorang. Kemampuan ini sangat erat berkaitan dengan anak sebagai individu yang mempunyai konsep diri, penghargaan terhadap diri sendiri (self esteem) dan mengatur diri sendiri (self regulation). Tujuan untuk meningkatkan daya cipta anak-anak dan memacu anak untuk belajar mengenal berbagai macam ilmu pengetahuan melalui pendekatan nilai budi bahasa, agama, sosial, emosional, fisik, motorik, kognitif, bahasa, seni dan kemandirian. Kemampuan anak dapat terlihat dari tingkah laku sehari-hari ataupun kebiasan yang dilakukan secara terus menerus, dalam proses ini menggunakan penilaian hasil evaluasi rapor selama 2 semester. Untuk mengetahui kemampuan dari masing-masing anak didiknya maka dibutuhkan sebuah sistem klasifikasi dengan proses text mining yaitu concept frequency-inverse document frequency (CF-IDF) merupakan proses analisis teks untuk menentukan nilai kecocokan antara dokumen pengetahuan dan keyword sedangkan untuk menghasilkan akurasi yang tepat menggunakan metode klasifikasi nearest neighbor (KNN) yang dilakukan pendekatan jarak antara masing masing objek dengan menggunakan jarak euclidean. Data yang digunakan merupakan hasil penilaian rapor selama 2 semester dengan jumlah 25 rapot dan kelas sebanyak 4 yaitu: sangat baik, baik, kurang dan sangat kurang. Hasil akurasi yang didapat dalam penelitian ini sebesar 50% menggunakan k-NN dengan nilai k=3.

Kata kunci: *Rapor, Kemampuan Anak, k-NN, Text Mining, Encludien*

1. Pendahuluan

Kemampuan merupakan kecakapan atau potensi seseorang untuk menguasai keahlian dalam melakukan sebuah pekerjaan yang beragam ataupun suatu penilaian atas tindakan seseorang. Kemampuan ini sangat erat berkaitan dengan anak sebagai individu yang mempunyai konsep diri, penghargaan terhadap diri sendiri (*self esteem*) dan mengatur diri sendiri (*self regulation*). Anak memahami tuntunan lingkungan terhadap dirinya sendiri dan penyesuaian tingkah lakunya. Di lihat kemampuan anak pada suatu kelas cenderung heterogen yang setiap kelasnya akan mengikuti gejala normal yang terdiri dari anak yang pandai, sedang dan kurang pandai.

Pendidikan anak di taman kanak-kanak ini memberikan rangsangan pendidikan untuk membantu pertumbuhan dan perkembangan jasmani dan rohani agar anak memiliki kesiapan dalam memasuki pendidikan yang lebih tinggi adapun tujuan untuk meningkatkan daya cipta anak-anak dan memacu anak untuk belajar mengenal berbagai macam ilmu pengetahuan melalui pendekatan nilai budi bahasa, agama, sosial, emosional, fisik, motorik, kognitif, bahasa, seni dan kemandirian. Kemampuan anak dapat terlihat dari tingkah laku sehari-hari ataupun kebiasan yang dilakukan secara terus menerus, dalam proses ini dilakukan dengan cara penilian hasil evaluasi rapor yang akan klasifikasi pola kemampuan dari masing-masing anak didik.

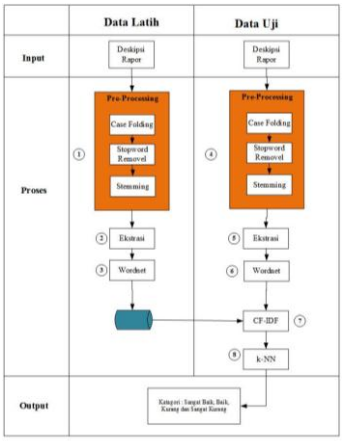
Penelitian terdahulu antara lain: Algoritma *k*-NN dapat digunakan dalam klasifikasi data Hasil Produksi Kelapa Sawit pada PT. Minamas Kec. Parindu. Berdasarkan hasil penelitian, data diklasifikasikan ke dalam 6 *cluster*. Berdasarkan hasil penelitian dapat dilihat kemiripan hasil produksi dari 50 kelompok tani yang ada di KUD. HIMADO. Nilai *k* yang di gunakan sebagai hasil pengamatan adalah *k*=7, karena untuk jarak minimum pada C1 memiliki persentase yang lebih besar yaitu 34%. Pada penelitian ini hasil produksi yang dominan adalah produksi dari kelompok tani kelapa sawit yang terletak pada C1. Dengan keanggotaan kelompok tani yaitu kelompok 1, 2, 33, 34, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 [5], k-NN dapat digunakan untuk menentukan kelayakan mobil menurut parameter kondisi fisik dari mobil tersebut, aplikasi data mining ini dapat memprediksi dengan menggunakan 1 data mobil atau 1 database dengan menggunakan data training yang berjumlah 14 data dengan jumlah k=3 didapat nilai accuracy 78%, sedangkan data training yang berjumlah

1728 data dengan k=11 didapat nilai accuracy 95.78%, nilai kappa statistic dan precission mendekati nilai 1, yang artinya bahwa metode KNN dapat digunakan untuk klasifikasi dengan memuaskan nilai ROC area juga mendekati 1 artinya sistim ini cukup akurat. Semakin besar jumlah data training sistem akan semakin akurat [3], perangkat lunak yang dapat membantu pihak administrasi jurusan untuk menentukan dosen mana yang sesuai untuk mengampu satu matakuliah dengan kompetensi yang dimilikinya. Untuk mendapatkan hasil yang maksimal, data dalam dokumen diisi sesuai dengan aturan yang ditentukan. Setiap kata yang digunakan dalam satu bidang ilmu yang mempunyai arti yanga sama diharapkan seluruhnya digunakan dan dimasukan dalam data yang telah ditentukan [1], Berdasarkan pengujian hasil prediksi menggunakan *algoritma k-nearest neighbor* secara manual dan menggunakan sistem yang digunakan data training adalah mengunakan 90 data mahasiswa yaitu 42 orang data teknik informatika S1, 40 orang mahasiswa sistem informasi S1 dan 8 orang mahasiswa teknik informatika D3, sistem didapatkan kesamaan hasil prediksi yaitu 79% dan melihat dari presentasi mungkin saja ini kurang akurat [6].

Berdasarkan hal diatas maka dalam klasifikasi pola kemampuan anak yang menggunakan hasil evaluasi rapor selama 2 semester untuk mengetahui kemampuan atau kelebihan dari masing-masing anak, maka pemelitian ini bertujuan dibutuhkan suatu sistem yang dapat mengklasifikasi pola kemampuan dari masing-masing anak didiknya.

1.1 Metode Penelitian

Proses awal pada sistem ini dengan input deskripsi penilaian rapor sebagai data latih yang sudah terdapat kelas dan data uji yang mewakili dari data uji yang ada, proses selanjutnya adalah *pre-processing* yang meliputi *case folding*, *stopword removal*, *stemming*, *ekstrasi* dan *wordnet*. Setelah melakukan pre-processing dilakukan perhitungan cf-idf dan setelah mendapatkan bobot yang sesuai dilakukan perhitungan *k-NN* untuk mendapatkan rangking dari setiap kategori antara lain Sangat Baik, Baik, Kurang dan Sangat Kurang. Ditunjukkan pada Gambar 1 merupakan metode penelitian.



Gambar 1. *Metode Penelitian*

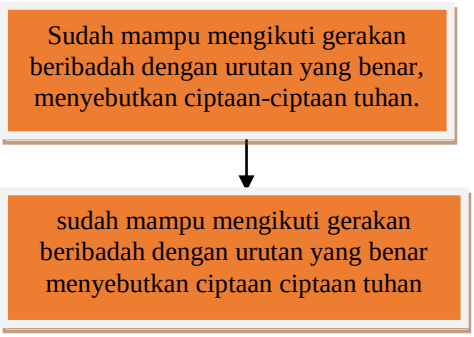
2. Pembahasan

2.1 Text Mining

Text mining adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen Word, PDF, kutipanteks. Jenis masukan untuk penambangan teks ini disebut data terstruktur dan merupakan pembeda utama dengan penambangan data yang menggunakan data terstruktur atau basis data sebagai masukan. Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevandari data teks terstrukturini dengan menggunakan teknik dan alat yang sama dengan penambangan data. Proses yang umum dilakukan oleh penambangan teks di antaranya adalah perangkuman otomatis, kategorisasi dokumen, penggugusan teks [1].

1. Case Folding

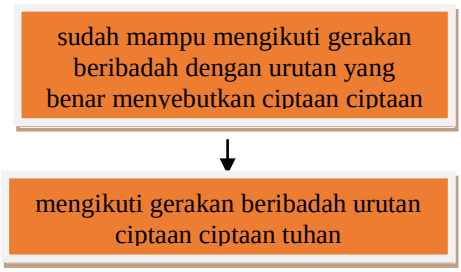
Tahap *case folding* mengubah huruf besar menjadi huruf kecil. Kemudian tanda baca titik pada akhir kalimat akan dihapuskan. Ditunjukkan pada Gambar 2 merupakan contoh *case folding*.



Gambar 2. *Contoh Case Folding*

2. Stopword Removal

Tahap *stopword removal* / *stoplist* yang merupakan kata penunjuk tempat akan dihapuskan, karena termasuk dalam kelompok kata tidak penting seperti "yang", "di", "saat", "jika" dan sebagainya. Ditunjukkan pada Gambar 3 merupakan contoh *stopword removal.*



Gambar 3. *Contoh Stopword Removal*

3. Stemming

Tahap *stemming* merupakan suatu proses untuk menemukan kata dasar dari sebuah kata agar sesuai dengan struktur *morfologi* bahasa Indonesia yang benar. Stemming dengan menghilangkan semua imbuhan baik terdiri awalan, sisipan, akhiran dan kombinasi awalan dan akhiran kata [1]. Tabel 1 ini menunjukkan tabel stemming kata yang digunakan.

Tabel 1. *Stemming Kata*

Kata	Hasil Stemming
Mengikuti	Ikuti
Gerakan	Gerak
Beribadah	Ibadah
Urutan	Urut
Ciptaan	Cipta
Ciptaan	Cipta
Tuhan	Tuhan

4. Mencari Konsep

Tahap ke-empat pada preprocessing melalui tahap ekstraksi. Tahap ini dilakukan untuk mencari *concept* dari setiap kata atau frase yang terdapat pada dokumen. *Concept* tersebut dapat berupa kata atau frase yang bersinonim atau pun memiliki makna yang sama. Untuk mendapatkan concept tersebut maka dilakukan pencarian dan pembentukan kandidat *concept* berdasarkan kedekatan kata (*adjacent*). Terdapat dua jenis kandidat concept yang akan dibangun yaitu:

1. Kandidat Kata (*mono word*) yaitu kandidat yang hanya terdiri dari satu kata.
2. Kandidat Frase (*multi words*) yaitu kandidat yang terdiri dari gabungan beberapa kata. Untuk pembentukan kandidat frase

Jumlah maksimum kata yang membentuknya dibatasi hanya tiga. Tabel 2 ini menunjukkan tabel contoh kandidat konsep.

Tabel 2. *Contoh Kandidat Konsep*

Kandidat Kata	Kandidat Frase
Ikuti	Ikuti gerak
Gerak	Gerak ibadah
Ibadah	Ibadah urut
urut	Urut cipta
Cipta	Cipta cipta
Cipta	Cipta tuhan
Tuhan	Tuhan ikuti gerak
	Ikuti gerak ibadah
	Gerak ibadah urut
	Ibadah urut cipta
	Urut cipta cipta
	Cipta cipta tuhan

Kandidat *concept* yang telah dibangun maka akan dipetakan ke dalam *wordnet* untuk dicari *concept*. *Concept* di dalam *wordnet* sendiri dibangun berdasarkan kesamaan makna dari kata atau frase. Hanya kandidat *concept* yang terdapat pada *wordnet* saja yang akan diperhitungkan [1]. Tabel 3 ini menunjukkan tabel contoh konsep yang digunakan.

Tabel 3. *Contoh Konsep*

Kata dan Frase	Concept
Ikuti	Ikuti
Gerak	Gerak
Ibadah	Ibadah
urut	urut
Cipta	Cipta
Cipta	Cipta
Tuhan	Tuhan
Ikuti gerak	Ikuti gerak
Gerak ibadah	Gerak ibadah
Ibadah urut	Ibadah urut
Urut cipta	Urut cipta
Cipta cipta	Cipta cipta
Cipta tuhan	Cipta tuhan
Tuhan ikuti gerak	Tuhan ikuti gerak
Ikuti gerak ibadah	Ikuti gerak ibadah
Gerak ibadah urut	Gerak ibadah urut
Ibadah urut cipta	Ibadah urut cipta
Urut cipta cipta	Urut cipta cipta
Cipta cipta tuhan	Cipta cipta tuhan

2.2 Concept Frequency – Inverse Document Frequency (CF-IDF)

Untuk menentukan nilai kecocokan antara dokumen pengetahuan dan *keyword* diperlukan pembobotan. Pembobotan atau disebut juga *weighting* merupakan pemberian bobot terhadap kata atau frase yang telah dihasilkan dari tahap sebelumnya. Model pembobotan tersebut dapat menggunakan pembobotan global, lokal atau kombinasi dari keduanya. Salah satu pembobotan

kombinasi tersebut adalah CF-IDF (*Concept Frequency-Inverse Document Frequency*). Pada metode ini tidak dilakukan perhitungan terhadap term (seperti pada TF-IDF) namun dengan menghitung *key concept* yang ditemukan dalam teks. Pada CF-IDF, dilakukan pendekatan representasi isi dokumen dengan menggunakan jaringan semantik yang disebut dokumen inti semantik. Dokumen tersebut kemudian dipetakan dalam jaringan semantik yang disebut *Wordnet* dan dikonversikan dari sekumpulan terms menjadi sekumpulan konsep (*concept*) [4]. Pendekatan ini membuat konsep dari CF-IDF terlihat lebih cerdas dibandingkan TF-IDF. *Concept* yang dimaksud dalam metode ini adalah kata atau pun istilah majemuk yang kombinasi katanya dapat memiliki banyak arti dan menimbulkan ambiguitas dalam pembacaannya. Untuk membentuk concept, terlebih dahulu harus dibentuk kandidat -kandidat concept dari dokumen. Kandidat-kandidat dibedakan menjadi kata (*mono word*) dan frase (*multi words*). Frase atau multi words merupakan gabungan dari beberapa kata yang memiliki arti. Pembentukan frase maksimal adalah terdiri dari gabungan tiga kata. Pembentukan kandidat kata berdasarkan kemunculan setiap kata di dalam dokumen sementara pembentukan kandidat frase dilakukan berdasarkan kedekatan kata berurutan dari kiri ke kanan [1].

$$cf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad \text{..... (1)}$$

Keterangan:

cf_{ij} = rasio frekuensi *concept* pada dokumen
$n_{i,j}$ = jumlah kemunculan *concept* dalam dokumen
$\sum_k n_{k,j}$ = total kemunculan seluruh *concept* dalam dokumen.

$$idf_i = \log \frac{[D]}{[\{d: c_i \epsilon d\}]} \quad \text{..... (2)}$$

Keterangan:

idf_i = rasio frekuensi dokumen
$[D]$ = jumlah total dokumen
$[\{d: c_i \epsilon d\}]$ = jumlah dokumen yang terdapat kemunculan concept.

$$w = cf_{i,j} * idf_i \quad \text{..... (3)}$$

Keterangan:

w = bobot CF-IDF
$cf_{i,j}$ = rasio frekuensi *concept* pada dokumen
idf_i = rasio frekuensi dokumen

No	Concept	Frequency					DF
		D1	D2	D3	…	D25	
1.	lagu	1	1	0	…	0	10
2.	islam	1	1	1	…	1	5
3.	Plastisin meronce	1	0	0	…	0	1
4.	meronce manik	0	0	1	…	0	7
5.	Mencampur warna finger	0	1	1	…	1	2
6.	Warna finger painting	0	0	0	…	1	1

1. Hitung nilai CF dari kata/konsep “lagu”
   a. $n_{i,j}$ = 1 dan $\sum_k n_{k,j}$ = 2 sehingga CFDQ = 1/2 = 0,5
2. Hitung *Inverse Document Frequency* (IDF) dari kata/konsep “lagu”:

$[D]$ = 25 dan df = 10 sehingga IDF = Log (25/10) = 2.5

3. Hitung Bobot CF-IDF dari kata “lagu”:
   CFDQ = 0,5 dan IDF = 2,5 sehingga CFIDF = 1,25

2.2 Klasifikasi Nearest Neighbor (k-NN)

Algoritma k-*Nearest Neighbor* (k-NN) merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. *k-NN* termasuk algoritma *supervised learning* dimana hasil *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada *k-NN*. Kelas yang paling banyak muncul itu yang akan menjadi kelas hasil klasifikasi [5]. Tujuan algoritma ini adalah adalah mengklasifikasikan objek baru berdasarkan atribut dan *training sample*. Algoritma k-NN menggunakan klasifikasi ketetanggaan (*neighbor*) sebagai nilai prediksi dari *query instance* yang baru. Algoritma ini sederhana, bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan ketetanggaannya. Langkah-langkah untuk menghitung algoritma k-NN antara lain [3] :

1. Menentukan parameter *k*
2. Menghitung jarak antara data yang akan dievaluasi dengan semua pelatihan.
3. Mengurutkan jarak yang terbentuk
4. Menentukan jarak terdekat sampai urutan *k*
5. Memasangkan kelas yang bersesuai
6. Mencari jumlah kelas tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi.

$$d_i = \sqrt{\sum_{i=1}^{p}(X_{1i} - X_{2i})^2} \quad \text{..... (4)}$$

Keterangan :

d = Jarak (*Distance*)

X_{1i} = Data Uji (*Testing*)

X_{2i} = Data Latih (*Training*)

P = Dimensi data

I = Variabel Data ke-

Pada penelitian ini menggunakan data rapor selama 2 semester. Penelitian ini memberi kategori SB: sangat baik, B: baik, K: kurang dan SK: sangat kurang. Tabel 4 ini menunjukkan table data latih yang digunakan.

Tabel 4. *Tabel Data Latih*

Nama	D1	D2	D3	D4	D5	D6	…	D25	Class
Jarrel	2,66	1,25	0	2,01	0	1,05	…	0	SB: Menari, B: Menggambar, K: Meronce, SK: Bernyanyi

Setelah ada data latih, maka kita perlu data uji untuk mengklasifikasi untuk mengetahui kemampuan anak. Tabel 5 ini memperlihatkan data uji sebagai berikut:

Tabel 5. *Data Uji*

Nama	D1	D2	D3	D4	D5	D6	…	D25	Class
Rahmi	1,01	0	0	0,90	1,15	2,00	…	0	?

Langkah terakhir yaitu menentukan nilai ketetanggaan menggunakan (k-NN) dengan menggunakan k = 3

$$U_1, L_1 = \sqrt{(1{,}01-2{,}66)^2 + (0-1{,}25)^2 + \ldots.} = 1{,}33$$

$$U_1, L_2 = \sqrt{(1{,}01-0)^2 + (0-0)^2 + \ldots.} = 0{,}98$$

……

……

$$U_1, L_{25} = \sqrt{(1{,}01-1{,}25)^2 + (0-2{,}01)^2 + \ldots.} = 3.04$$

pada proses *k-NN* yaitu menggunakan k=3. Dengan menghitung jarak kedekatan menggunakan rumus *euclidean distance*. Tabel 6 ini hasil data *Euclidean* sebagai berikut:

Tabel 6. *Hasil Nilai k-NN*

D1	D2	D4	D25
1,33	0,98	3,00	3,04

Kategori pada dokumen :

D1 = Sangat Baik

D2 = Baik

D4 = Kurang

D25 = Sangat Kurang

Dokumen uji termasuk kategori sangat baik menggunakan parameter k=3

3. Kesimpulan

Uji Akurasi digunakan untuk mengetahui tingkat akurasi dari algoritma *k-Nearest Neighbor* yang didapat dari proses *CF-IDF*. Sebelumnya terdapat 25 data rapor, untuk uji akurasi data latih yang digunakan 25 data dan diketahui bahwa k=3 mendapatkan hasil 15 data yang sama dengan data real. Dari hasil tersebut diketahui bahwa k=3 mendapatkan hasil akurasi sebesar 50%.

Daftar Pustaka

[1] Y. H. Pudjiantoro, "Analisa Kompetensi Dosen Dalam Menentuan Matakuliah Yang Diampu Menggunakan Metode CF-IDF," *Aristoteles*, Vols. Vol.10, No. 1, pp. 1-8, Oktober 2012.

[2] R. K. Hapsari and Y. J. Santoso, "Stemming Artikel Berbahasa Indonesia Dengan Pendekatan Confix-Stripping," *Prosiding Seminar Nasional Manajemen Teknologi XXII*, Vols. ISBN : 978-602-10604-1-8, pp. 1-8, 24 Januari 2015.

[3] A. Nouvel, "Klasifikasi Kendaraan Roda Empat Berbasis K-NN," *Jurnal Blanglala Informatika*, Vols. Vol 3, No 2, pp. 66-69, September 2015.

[4] T. A. Hermawan, Y. H. Chrisnanto and A. I. Hadiana, "Klasifikasi Helpdesk Universitas Jenderal Achmad Yani Menggunakan CF-IDF dan K-NN," *Prosiding SNST ke-7*, Vols. ISBN : 978-602-99334-5-1, pp. 108-113, 2016.

[5] N. Krisandi, Helmi and B. Prihandono, "Algoritma k-Nearest Neighbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada PT. MINAMAS Kecamatan Parindu," *Buletin Ilmiah Math. Stat. dan Terapan (Bimaster)*, Vols. Vol 02, No. 1, pp. 33-38, 2013.

[6] R. I. Ndamanu, Kursini and M. R. Aif, "Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa Dengan Metode K-Nearest Neighbor," *Jatisi*, Vols. Vol 1, No. 1, pp. 1-13, September 2014.

Biodata Penulis

Putri Eka Prakasawati, sedang menempuh pendidikan sarjana (S1) jurusan Program Studi Informatika di Universitas Jenderal Achmad Yani Cimahi

Gunawan Abdillah, memperoleh gelar Sarjana (S.Si), Jurusan Universitas Jenderal Achmad Yani Cimahi, lulus tahun 2001. Memperoleh gelar Magister Komputer (M.Cs) Program Pasca Sarjana Magister Informatika Universitas Gajah Mada Yogyakarta, lulus tahun 2009. Saat ini menjadi Dosen di Universitas Jenderal Achmad Yani Cimahi.

Asep Id Hadiana, memperoleh gelar Sarjana (S.Si), Jurusan Ilmu Komputer Universitas Pajajaran Bandung, lulus tahun 2002. Memperoleh gelar Magister Ilmu Komputer (M.Kom) Program Pasca Sarjana Magister Ilmu Komputer Universitas Komputer Indonesia Bandung, lulus tahun 2010. Saat ini menjadi Dosen di Universitas Jenderal Achmad Yani Cimahi.

ISSN : 2302-3805