Implementasi Algoritma K-Nearest Neighbor Berbasis Forward Selection untuk Prediksi Mahasiswa non Aktif Universitas Dian Nuswantoro Semarang

Abu Salam[1], Ferry Bintang Nugroho[2], Junta Zeniarja[3]
Program Studi Teknik Informatika – S1, Universitas Dian Nuswantoro
Jl. Nakula I No 5 – 7 Semarang, 50131, Telp (024)3515261
E-mail: [1]abu.salam@dsn.dinus.ac.id , [2]111201508782@mhs.dinus.ac.id, [3]junta@dsn.dinus.ac.id

Diterima: 11 Februari 2020; Direvisi: 23 April 2020; Disetujui: 6 Mei 2020

Abstrak

Masalah yang muncul berkaitan dengan status mahasiswa salah satunya adalah status mahasiswa yang non aktif. Beberapa faktor penyebab status non aktif tersebut diantaranya adalah faktor ekonomi, kemampuan akademik, dan lain – lain. Manajemen perguruan tinggi perlu mengidentifikasi serta melakukan tindakan terhadap mahasiswa yang mempunyai status "tidak diharapkan" untuk mengetahui faktor munculnya masalah tersebut perlu dilakukan evaluasi saat pertengahan masa studi mahasiswa guna mencegah sedini mungkin munculnya mahasiswa yang diindikasi terdapat status tidak aktif untuk mengurangi dampak yang ditimbulkan akibat status non aktif tersebut. Penelitian Prediksi mahasiswa non aktif menggunakan algoritma klasifikasi K-Nearest Neighbor yang dikombinasikan dengan metode forward selection untuk seleksi atribut terbukti mampu meningkatkan nilai akurasi pada proses klasifikasi. Nilai akurasi yang didapatkan pada algoritma K-Nearest Neighbor tanpa forward selection sebesar 96.43% sedangkan pada algoritma K-Nearest Neighbor berbasis Forward Selection meningkat menjadi 97.27%.

Kata Kunci: *mahasiswa non aktif, forward selection, k-nearest neighbor*

Abstract

Problems that arise relating to the status of students one of which is the status of students who are not active. Some factors causing the non-active status include economic factors, academic ability, and others. Higher education management needs to identify and take action against students who have the status of "unexpected" to find out the factors that arise the problem needs to be evaluated during the middle of the student study period to prevent as early as possible the emergence of students indicated there is an inactive status to reduce the impact caused by the status inactive. In this study prediction of non-active students have been done using the K-Nearest Neighbor classification algorithm combined with the forward selection method for attribute selection is increase the accuracy value in the classification process. The accuracy value obtained on the K-Nearest Neighbor algorithm without Forward Selection is 96.43% while the K-Nearest Neighbor algorithm based on Forward Selection is 97.27%.

Keywords: *not active student, forward selection, k-nearest neighbor*

1. PENDAHULUAN

Masalah yang muncul berkaitan dengan status mahasiswa salah satunya adalah status

mahasiswa yang non aktif. Berdasarkan data dari Pusat Data dan Informasi Kemenristek Dikti Republik Indonesia [1] bahwa pada tahun akademik 2012/2013 sampai tahun 2016/2017 jumlah rata - rata mahasiswa baru yang diterima di seluruh perguruan tinggi di Indonesia berjumlah 1.322.576 sedangkan lulusan pada tahun akademik 2012/2013 sampai tahun 2016/2017 rata - rata berjumlah 915.347. Jumlah lulusan tersebut mencapai 69,20% dari rata - rata jumlah mahasiswa baru setiap tahunnya. Dengan kata lain terdapat sekitar 30,80% mahasiswa yang tidak jelas status yang dimilikinya. Status yang tidak jelas tersebut dapat diasumsikan karena mahasiswa memiliki status mangkir (non aktif), menempuh studi tidak tepat waktu, atau drop out. Pada tahun 2017 saja terdapat 195.176 kasus mahasiswa yang drop out atau sekitar 2,8% di seluruh perguruan tinggi di Indonesia.

Beberapa faktor penyebab status yang tidak jelas tersebut diantaranya adalah faktor ekonomi, kemampuan akademik, dan lain – lain. Manajemen perguruan tinggi perlu mengidentifikasi serta melakukan tindakan terhadap mahasiswa yang mempunyai status "tidak diharapkan" untuk mengetahui faktor munculnya masalah tersebut perlu dilakukan evaluasi saat pertengahan masa studi mahasiswa guna mencegah sedini mungkin munculnya mahasiswa yang diindikasi terdapat status tidak aktif untuk mengurangi dampak yang ditimbulkan.

Salah satu langkah yang mampu dijalankan dalam mengidentifikasi permasalahan tersebut yaitu dengan melakukan analisis pola atau serangkaian informasi yang didapat dari pusat data perguruan tinggi mengenai faktor - faktor yang mempengaruhi munculnya mahasiswa berstatus non aktif menggunakan metode data mining. Pada penelitian sebelumnya yang dijalankan oleh Khafiizh Hastuti [2] tahun 2012 menunjukkan hasil bahwa metode decision tree mempunyai hasil yang paling akurat dibandingkan dengan algoritma klasifikasi lainnya yaitu sebesar 95.29% untuk memprediksi mahasiswa non aktif. Penelitian yang akan dilakukan ini merupakan penelitian perbandingan penggunaan algoritma serta berfokus pada teknik klasifikasi data *mining* untuk mendapatkan nilai akurasi yang lebih tinggi dengan menggunakan sampel dari dataset mahasiswa Program Studi Teknik Informatika Universitas Dian Nuswantoro Semarang.

Berikutnya Penelitian yang dijalankan oleh Yeyen D. Atma, dan Arif [3] membuktikan bahwa penggunaan algoritma *K-Nearest Neighbor* dengan kombinasi *forward selection* untuk mengidentifikasi mahasiswa berpotensi *drop out* memiliki nilai akurasi yang sangat tinggi yaitu sebesar 99.46%. Sedangkan penelitian yang dilakukan oleh Eri S. Susanto, Kusrini dan Hanif Al Fatta [4] untuk memprediksi kelulusan mahasiswa menggunakan algoritma KNN mendapatkan hasil nilai akurasi sebesar 98.46%

2. METODE PENELITIAN

Teknologi data mining adalah metode yang berasal dari berbagai macam bidang disiplin ilmu, namun bidang ilmu yang paling mendasar adalah *machine learning*, statistika, matematika, dan *artificial intelligence*. Teknik ini banyak digunakan untuk mengidentifikasi, menemukan serta menguraikan pola informasi atau pengetahuan yang bermanfaat di pusat data [5].

Untuk mendukung penelitian ini penulis memakai salah satu standar pemrosesan data mining yaitu CRISP-DM dimana siklus hidup tersebut dibagi menjadi enam tahapan atau fase [6] yang meliputi pemahaman bisnis, pemahaman data, *data prepraration*, yang kemudian dilakukan proses mining untuk mendapatkan hasil akhir dari penelitian ini. Pada pemodelan ini mengimplementasikan dua metode yang akan dipakai yaitu *K-Nearest Neighbor* dengan *Forward Selection* dan *K-Nearest Neighbor*, berdasarkan hasil dari pemodelan tersebut dilakukan uji perhitungan nilai akurasi menggunakan metode *Confusion Matrix.*

Mahasiswa Non Aktif merupakan mahasiswa yang tidak melakukan aktifitas registrasi akademik [7]. Registrasi akademik adalah kegiatan mendaftarkan diri menjadi mahasiswa peserta kegiatan akademik tiap semester berupa pengisian Kartu Rencana Studi.

Teknik *Forward Selection* merupakan metode pemodelan yang dimulai dari *empty model*, selanjutnya satu per satu peubah dimasukkan hingga kriteria tertentu terpenuhi. Penggunaan teknik *forward selection* yaitu untuk meningkatkan hasil akurasi fitur yang tidak relevan harus

dihapus agar tidak mengurangi nilai akurasi dan menyebabkan gangguan [8]. Berikut merupakan langkah – langkah dari *forward selection* [3]:

a. Membentuk model dengan cara meregresikan variabel respon Y dengan tiap variabel predictor. Lalu pilih model yang memiliki nilai R^2 yang paling tinggi. Misal model tersebut mempunyai prediktor X_a, yaitu pada persamaan (1) :

$$\hat{Y} = b_0 + b_a X_a \quad (1)$$

b. Kemudian meregresikan variabel respon Y, dengan prediktor X_a, dijumlah dengan setiap prediktor selain dari X_a serta prediktor yang lain. Setelah itu tentukan model dengan nilai R^2 yang paling tinggi, misal memiliki tambahan prediktor X_b, misal model pada persamaan (2) :

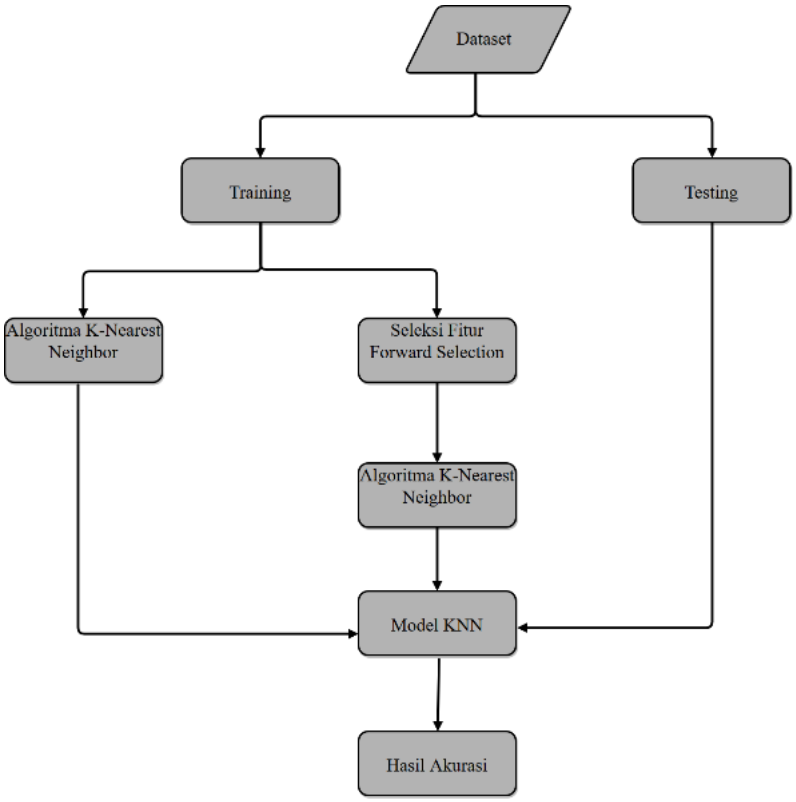
$$\hat{Y} = b_0 + b_a X_a + b_x X_b \quad (2)$$

c. Prediktor terpilih X_b artinya memiliki Fsequensial yang paling tinggi. Formula $F_{sequensial}$ untuk X_b adalah

$$F_{seq} = R(\beta_b \beta_0 \beta_a) MSE\ db. \quad (3)$$

Dimana pada persamaan (3) nilai $F_{sequensial}$ untuk X_b juga bisa didapat dengan cara mengkuadratkan statistik uji T prediktor X_b.

d. Proses tersebut diulang hingga didapatkan $F_{sequensial} > F_{in}$.
Nilai dari $F_{in} = F(1, v, \alpha_{in})$, jadi model terbaik yang didapatkan adalah model yang tidak memiliki prediktor dengan $F_{sequensial} < F_{in}$



Gambar 1. Fase Pemodelan

Algoritma *K-Nearest Neighbor* yaitu suatu teknik / metode yang biasa digunakan untuk mengklasifikasi suatu data atau teks. Algoritma ini memiliki fungsi dengan mengelompokkan data yang baru berdasarkan jarak data baru tersebut ke dalam beberapa data tetangga (*neighbor*)

yang terdekat.

Cara kerja yang dilakukan oleh KNN adalah untuk memperoleh class kategori melalui proses menghitung antara data baru dengan tiap data yang sudah dikategorikan sebelumnya. Pencarian jarak terdekat tersebut dibutuhkan untuk menghitung jumlah kemiripan antar data yang sudah diproses [9]. Istilah tersebut sering juga disebut dengan *supervised learning*, yaitu suatu metode pembelajaran terawasi dimana hasil yang diperoleh sudah diketahui sebelumnya dari data yang sudah ada.

Agar lebih jelas berikut merupakan tahap – tahap yang harus dilakukan di dalam algoritma *K-Nearest Neighbor* [10]:

1. Menentukan parameter k (*neighbord*) yang akan digunakan.
2. Menghitung jarak antara data uji dengan data latih.
3. Lakukan sorting atau pengurutan dari jarak yang sudah terbentuk.
4. Pilih jarak yang terdekat sesuai dari parameter k yang diambil
5. Pasangkan *class* yang sesuai.
6. Hitung jumlah class dari kategori mayoritas tetangga terdekat dan gunakan sebagai prediksi nilai yang baru.

Tujuan dari algoritma K-NN yaitu untuk klasifikasi data berdasarkan variable serta sampel data training. Klasifikasi tersebut menggunakan voting yang paling banyak antara klasifikasi objek k yang sudah ditentukan. Dalam menentukan pencarian jarak terdekat akan digunakan perhitungan *Euclidean Distance*. Untuk perhitungan rumus dari *K-Nearest Neighbor* dapat dijelaskan pada persamaan (4) sebagai berikut :

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

Dimana p = Dimensi data; i = Atribut data; x_1 = Data sampel / data latih; x_2 = Data uji / data testing; d = Jarak

Untuk mengetahui performa algoritma yang digunakan salah satu metode yang dapat digunakan adalah *Confusion Matrix* pada tabel 1 [11]. Metode tersebut mengandung informasi paling aktual tentang perkiraan klasifikasi yang dihasilkan sistem dengan hasilnya menggunakan data ke dalam bentuk sebuah matriks. Berikut merupakan penjelasan dari tabel *confusion matrix*.

Tabel 1. Confusion Matrix

Classification (class)	Predicted Class	
	Yes	No
YES	a (True Positive - TP)	b (False Negative - FN)
NO	c (False Positive - FP)	d (True Negative - TN)

Accuracy pada perasmaan (5) yaitu jumlah proporsi banyaknya prediksi yang bernilai benar.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

True Positive (TP), yaitu jumlah klasifikasi tidak normal yang mempunyai nilai benar yang biasa dikenal sebagai *sensitivity*. *False Positive* (FP), yaitu jumlah kasus normal yang kesalahannya diklasifikasikan sebagai kelas yang tidak normal. *False Negative* (FN), yaitu jumlah kasus normal yang kesalahannya diklasifikasikan sebagai kelas yang normal. *True Negative* (TN), yaitu jumlah proporsi benar yang mempunyai nilai negatif yang biasa dikenal sebagai *specificity*.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini digunakan dataset mahasiswa Program Studi Teknik Informatika – S1 tahun angkatan 2014 – 2016 pada tahun akademik 2018/2019 Genap seperti pada tabel 2, kemudian hasil yang diperoleh berupa nilai akurasi dari proses pengujian yang dilakukan yaitu dari algoritma *K-Nearest Neighbor* dengan kombinasi *Forward Selection*. Teknik seleksi fitur *forward selection* yang dipakai bertujuan agar menghilangkan variabel – variabel yang kurang berpengaruh saat proses klasifikasi. Berikut merupakan atribut yang digunakan pada penelitian ini:

Tabel 2. Deskripsi Atribut Dataset

No.	Atribut	Tipe Data	Keterangan
1	Status Menikah	Binominal	Status mahasiswa apakah sudah menikah atau belum
2	Status Kerja	Binominal	Status mahasiswa tersebut apakah sudah bekerja atau belum pada saat menjalani studinya
3	Beasiswa	Binominal	Apakah mahasiswa yang bersangkutan merupakan penerima beasiswa atau tidak
4	Jalur Pendaftaran	Binominal	Adalah jalur pendaftaran yang diambil oleh mahasiswa yang bersangkutan ketika mendaftarkan diri ke Universitas Dian Nuswantoro Semarang meliputi jalur PMDK atau Reguler
5	Gaji Orangtua	Polinominal	Merupakan total gaji orangtua dari mahasiswa yang dikategorikan menjadi 6 kategori yaitu < 3 juta, 3 juta – 5 juta, 5 juta – 7 juta, 7 juta – 10 juta, > 10 juta, dan tidak diketahui
6	Kota Asal	Polinominal	Yaitu kota asal mahasiswa bersangkutan yang dibagi menjadi 3 kategori yaitu Semarang, Non Semarang, dan tidak diketahui
7	Umur saat mendaftar	Integer	Adalah umur mahasiswa disaat mendaftar perkuliahan di Universitas Dian Nuswantoro
8	SKS 1	Integer	Yaitu total jumlah Satuan Kredit Semester yang diambil saat semester 1
9	SKS 2	Integer	Yaitu total jumlah Satuan Kredit Semester yang diambil saat semester 2
10	SKS 3	Integer	Yaitu total jumlah Satuan Kredit Semester yang diambil saat semester 3
11	SKS 4	Integer	Yaitu total jumlah Satuan Kredit Semester yang diambil saat semester 4
12	IPS 1	Real	Yaitu Indeks Prestasi Semester yang diperoleh saat semester 1
13	IPS 2	Real	Yaitu Indeks Prestasi Semester yang diperoleh saat semester 2
14	IPS 3	Real	Yaitu Indeks Prestasi Semester yang diperoleh saat semester 3
15	IPS 4	Real	Yaitu Indeks Prestasi Semester yang diperoleh saat semester 4
16	Status Mahasiswa	Binominal	Status mahasiswa tersebut apakah aktif atau non aktif yang merupakan sebagai *class*/label

Untuk mengetahui tingkat akurasi yang didapatkan pada proses klasifikasi menggunakan algoritma *K-Nearest Neighbor* tanpa seleksi fitur *forward selection* tersebut maka akan digunakan nilai k=3, k=5, k=7, k=9. Berikut merupakan hasil nilai akurasi yang didapatkan :

Tabel 3. Hasil Akurasi Tiap Nilai k Algoritma KNN

K	Akurasi
3	96.43%
5	91.95%
7	91.11%
9	87.67%

Berdasarkan tabel 3 diatas nilai k yang mempunyai tingkat akurasi yang paling tinggi adalah k=3 dengan nilai akurasi yaitu sebesar 96.43% dilanjutkan dengan k=5 sebesar 91.95%, k=7 sebesar 91.11%, dan k=9 sebesar 87.67%. Berikut merupakan diagram *confusion matrix* pada gambar 2 untuk nilai k=3 :

accuracy: 96.43% +/- 1.58% (mikro: 96.43%)

	true Non Aktif	true Aktif	class precision
pred. Non Aktif	380	13	96.69%
pred. Aktif	38	997	96.33%
class recall	90.91%	98.71%	

Gambar 2. Hasil Nilai Akurasi Algoritma KNN Untuk K=3

$$Akurasi = \frac{997 + 380}{997 + 380 + 13 + 38} x100\% = 96.43\%$$

3.1 Algoritma *K-Nearest Neighbor* Berbasis *Forward Selection*

Hasil yang didapat pada perhitungan *forward selection* dapat diketahui bahwa didapatkan 6 atribut yang memiliki hubungan kuat terhadap atribut target, yaitu Status Menikah, Kota Asal, SKS 4, IPS 1, IPS 3, dan IPS 4.

Dari variabel – variabel yang telah dilakukan seleksi fitur menggunakan forward selection kemudian dilakukan proses klasifikasi memakai algoritma *K-Nearest Neighbor* untuk klasifikasi mahasiswa non aktif. Dalam penelitian yang dilakukan ini akan dilakukan perbandingan akurasi yang didapatkan dalam proses klasifikasi menggunakan nilai k=3, k=5, k=7, dan k=9. Berikut pada tabel 4 merupakan hasil yang diperoleh dari nilai k yang sudah ditetapkan:

Tabel 4. Hasil Akurasi Tiap Nilai K Algoritma KNN Berbasis Forward Selection

K	Akurasi
3	97.27%
5	94.33%
7	92.65%
9	89.91%

Dilakukan tahap pengujian menggunakan diagram confusion matrix untuk mendapatkan hasil akurasi yang didapatkan menggunakan algoritma *K-Nearest Neighbor* berbasis *Forward Selection* dengan k=3 dan hasil ditampilkan pada gambar 3:

accuracy: 97.27% +/- 0.97% (mikro: 97.27%)

	true Non Aktif	true Aktif	class precision
pred. Non Aktif	387	8	97.97%
pred. Aktif	31	1002	97.00%
class recall	92.58%	99.21%	

Gambar 3. Hasil Nilai Akurasi Algoritma KNN Berbasis Forward Selection Untuk K=3

$$Akurasi = \frac{1002 + 387}{1002 + 387 + 8 + 31} x100\% = 97.27\%$$

Berdasarkan hasil penelitian proses prediksi mahasiswa non aktif dengan Algoritma KNN menggunakan forward selection nilai akurasi maksimal yang didapatkan sebesar 97.27%, nilai akurasi tersebut merupakan nilai akurasi yang sangat baik pada saat dilakukan deployment, sehingga Universitas Dian Nuswantoro kedepan dapat melakukan proses prediksi mahasiswa non aktif dengan lebih akurat, dan dapat dijadikan bahan evaluasi untuk dilakukan penanganan lebih dini.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan didapatkan beberapa kesimpulan berdasarkan hasil implementasi algoritma *K-Nearest Neighbor* berbasis *Forward Selection* dalam mengklasifikasikan mahasiswa non aktif yaitu bahwa penggunaan teknik *Forward Selection* untuk seleksi fitur terbukti mampu meningkatkan nilai akurasi dari algoritma *K-Nearest Neighbor*, dimana pada awalnya atribut yang digunakan berjumlah 15 berkurang menjadi 6 atribut yang dipakai untuk klasifikasi status mahasiswa non aktif. Hasil proses pengujian algoritma *K-Nearest Neighbor* tanpa menggunakan seleksi fitur *Forward Selection* yaitu sebesar 96.43%, pada pengujian algoritma *K-Nearest Neighbor* berbasis *Forward Selection* terjadi peningkatan hasil akurasi menjadi 97.27%. Dari hasil penelitian ini dapat dibuktikan bahwa algoritma *K-Nearest Neighbor* dengan kombinasi teknik *Forward Selection* pada kasus klasifikasi mahasiswa non aktif lebih baik dibandingkan dengan algoritma *K-Nearest Neighbor* saja. Nilai akurasi yang didapatkan sebesar 97.27% merupakan nilai akurasi yang sangat baik untuk dasar deployment aplikasi proses prediksi mahasiswa non aktif kedepannya.

5. SARAN

Pada penelitian ini terdapat beberapa saran yang perlu diperhatikan agar dapat melakukan selanjutnya untuk mendapatkan hasil yang lebih baik yaitu perlu dilakukan penelitian menggunakan teknik seleksi fitur yang lain seperti *backward elimination* atau yang lainnya yang mungkin mampu menghasilkan nilai akurasi yang lebih baik untuk penelitian selanjutnya. Kemudian dilakukan pencatatan serta penyimpanan status mahasiswa tiap semesternya agar dapat diketahui berapa lama mahasiswa tersebut non aktif serta faktor apa saja yang mempengaruhinya. Dan yang terakhir perlu dilakukan penelitian lanjutan untuk proses deployment aplikasi yang dapat digunakan secara langsung oleh Biro Kemahasiswaan Universitas Dian Nuswantoro untuk secara periodik dapat melakukan prediksi mahasiswa non aktif.

DAFTAR PUSTAKA

[1] Tim Penyusun Statistik Pendidikan Tinggi Tahun 2017, Statistik Pendidikan Tinggi Tahun 2017, Kemenristekdikti Republik Indonesia, 2017.

[2] K. Hastuti, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif," in *Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012*, Semarang, 2012.

[3] Y. Atma and A. Setyanto, "Perbandingan Algoritma C4.5 dan K-NN Dalam Identifikasi Mahasiswa Berpotensi Drop Out," *Metik Jurnal*, vol. II, pp. 31-37, 2018.

[4] Kusrini, E. Sasmita and H. Al Fatta, "Prediksi Kelulusan Mahasiswa Magister Teknik Informatika Universitas Amikom Yogyakarta Menggunakan Metode K-Nearest Neighbor," *Jurnal Teknologi Informasi*, vol. XIII, pp. 67-72, 2018.

[5] D. Nofriansyah, Konsep Data Mining Vs Sistem Pendukung Keputusan, Yogyakarta: Deepublish, 2014.

[6] D. Larose and C. Larose, Data Mining and Predictive Analytics (Second Edition), New Jersey: John Wiley & Sons Inc, 2015.

[7] Tim Penyusun Keputusan Rektor UDINUS, "Keputusan Rektor Universitas Dian Nuswantoro Nomor : 077/KEP/UDN-01/VIII/2018 Tentang Peraturan Akademik Universitas Dian Nuswantoro," Universitas Dian Nuswantoro, Semarang, 2018.

[8] S. Tabakhi, P. Moradi and F. Akhlaghian, "An unsupervised feature selection algorithm based," *Engineering Applications of Artificial Intelligence*, vol. XXXII, pp. 112-123, 2014.

[9] E. Hardiyanto and F. Rahutomo, "Studi Awal Klasifikasi Artikel Wikipedia Bahasa Indonesia Dengan Menggunakan Metoda K Nearest Neighbor," in *Seminar Nasional Terapan Riset Inovatif*, Semarang, 2016.

[10] M. Lestari, "Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) Untuk Mendeteksi Penyakit Jantung," *Faktor Exacta*, vol. VII, no. 4, pp. 366-371, 2014.

[11] C. Vercellis, Business Intelligence:Data Mining and Optimizationfor Decision Making, Milan: John Wiley & Sons Ltd, 2009.