

Using Fuzzy K-Nearest Neighbor for Predicting University Students Graduation In Teknokrat

¹Imam Ahmad, ²Heni Sulistiyani, ³Hendrik Saputra

^{1,2,3}Sistem Informasi, Fakultas Teknik Dan Ilmu Komputer, Universitas Teknokrat Indonesia

Email: 1imamahmad@teknokrat.ac.id, 2@gmail.com, 3hendriksa68@yahoo.com

Article Info

Article history:

Received Feb 12th, 2018

Revised Feb 30th, 2018

Accepted Mar 27th, 2018

Keywords:

Fuzzy

K-Nearest Neighbor

Prediction

Student Graduation Rate

ABSTRACT

The absence of prediction system that can provide prediction analysis on the graduation rate of students becomes the reason for the research on the prediction of the level of graduation rate of students. The number of students who graduate on time becomes indicator of success of universities. Many studies on predicting students on time graduation have been conducted. The Fuzzy K-Nearest Neighbor, a method to classify objects based on training data located closest to the object, and using data from teknokrat universities student graduates in 2009 until 2013 faculties of engineering and computer science was used in this study. Selection of the value of k in KNN algorithm became important because it would affect the performance of the algorithm KNN. Based on the results of measuring the performance of the algorithm by using cross validation, confusion matrix, it was found the value of k = 10 and the fuzzy K-Nearest Neighbor had an accuracy value of 77,35%.

Copyright © 2018 Puzzle Research Data Technology

Corresponding Author:

Imam Ahmad

Sistem Informasi, Fakultas Teknik dan Ilmu Komputer

Universitas Teknokrat Indonesia

Jl. Zaenal Abidin Pagar Alam No. 9-11, Labuhan Ratu, Kota Bandar Lampung 35132, Lampung

Email: imamahmad@teknokrat.ac.id

1. PENDAHULUAN

Data Mining didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, data mining dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering dan asosiasi.[6] Selain itu data mining merupakan salah satu teknik yang dapat melakukan prediksi dengan teknik penggalian data. Penggalian data berdasarkan data pendidikan di universitas dapat meningkatkan kualitas pembelajaran Mahasiswa di universitas.

Melakukan Prediksi Tentang prediksi tingkat kelulusan Mahasiswa sangatlah penting dimana informasi yang dihasilkan oleh hasil prediksi ini dapat dijadikan sebagai bahan pertimbangan bagi pihak manajemen program studi untuk melakukan langkah secara persuasif dalam rangka meningkatkan persentase kelulusan Mahasiswa yang memenuhi standar mutu yang telah ditetapkan kampus. Prediksi dalam bentuk komputasi merupakan satu kegiatan matematis. Prediksi ini telah dilakukan bertahun-tahun sebelum penguasaan komputer, yaitu menggunakan kalkulator. Prediksi tingkat kelulusan Mahasiswa dapat membantu pihak manajemen dalam mengambil keputusan. Bagaimana menganalisa kelulusan Mahasiswa dan mendapatkan parameter terbaik dari metode yang digunakan untuk dapat memprediksi kelulusan Mahasiswa.

Sebuah penelitian membahas tentang Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa dengan Metode K-Nearest Neighbor. Dimana dalam penelitian yang dilakukan oleh penulis mengangkat masalah mengenai pentingnya analisis prediksi tentang pengunduran diri Mahasiswa di STIKOM UYELINDO Kupang karena menentukan prediksi tingkat pengunduran diri Mahasiswa dalam jumlah besar tidak mungkin dilakukan secara manual karena membutuhkan waktu yang cukup lama. Untuk itu dibutuhkan sebuah algoritma yang dapat mengkategorisasikan prediksi tingkat pengunduran diri Mahasiswa secara otomatis menggunakan komputer. Hasil yang didapat dari penelitian ini yaitu bertujuan untuk menghasilkan sebuah data yang dapat diuji untuk mendapat prediksi dan keakuratan klasifikasi [8].

Adapun penelitian lainnya membahas tentang Graduation Prediction of Gunadarma University Students Using Algorithm and Naïve Bayes C4.5 Algorithm, Permasalahan yang dihadapi adalah banyaknya Mahasiswa yang lulus tidak tepat waktu. Untuk mengetahui tingkat kelulusan Mahasiswa dalam satu tahun ajaran dapat dilakukan suatu prediksi berdasarkan data-data Mahasiswa pada tahun ajaran pertama. Algoritma yang digunakan adalah C45 dan naïve bayes. Hasil dari penelitian ini adalah akurasi untuk metode naïve bayes adalah 80,85% dengan presentasi kesalahan 19,05% Akurasi ketepatan hasil prediksi C4.5 85,7% dan presentasi kesalahannya adalah 14,3% [11].

Penelitian yang dilakukan oleh Karamouiz [12] dengan judul Sensitivity Analysis of Neural Network for Identifying the Factors for College Students Success, Masalah yang dikaji adalah tingkat kelulusan yang dianggap sebagai indikator efektivitas suatu lembaga institusi, Metode yang digunakan yaitu Neural Network (NN). Dari hasil data training yang dilakukan, diperoleh kategori yang lulus adalah 86,04% dan data training yang tidak sukses adalah 68,21 % dan error yang diperoleh untuk kedua kategori tersebut adalah 0,18%.

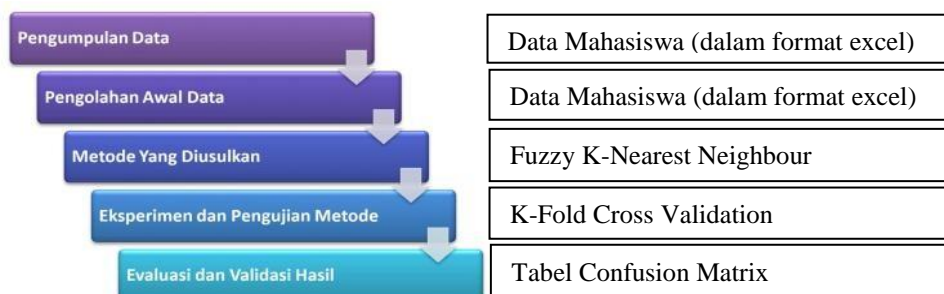
Penelitian yang dilakukan oleh Qudri, Kalyankar (2010) dengan judul Drop Out Feature of Student Data for Academic Performance Using Decision Tree techniques. Masalah dalam penelitiannya adalah prestasi akademik siswa sangat penting bagi lembaga pendidikan karena program-program strategis dapat direncanakan untuk meningkatkan atau mempertahankan prestasi siswa selama periode mereka studi di lembaga. Metode yang digunakan adalah Decision Tree, yakni algoritma J4.8. Hasil dari penelitian ini adalah sebuah pohon keputusan yang dapat dijadikan rule bagi prediksi siswa yang putus sekolah.[9]

Penelitian yang dilakukan oleh Mustakim, Giantika Oktaviani F (2016) Program Studi Sistem Informasi Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim, Riau dengan judul Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa. Dimana dalam penelitian yang dilakukan oleh penulis mengangkat masalah tentang Persoalan Indeks Prestasi Mahasiswa. Atribut yang digunakan dalam proses prediksi adalah Jenis Kelamin, Jenis Tinggal, Umur, Jumlah Satuan Kredit Semester (SKS), dan Jumlah Nilai Mutu (NM), sehingga dengan menerapkan algoritma K-Nearest Neighbor dapat dilakukan prediksi berdasarkan kedekatan dari histori data lama (training) dengan data baru (testing). Hasil yang didapat dari penelitian ini adalah sistem prediksi yang dapat dijadikan sebagai acuan bagi Mahasiswa untuk meningkatkan prestasi dan predikat perkuliahan dimasa yang akan datang [7].

Adapun yang membedakan penelitian yang dilakukan oleh penulis sebelumnya yaitu penelitian ini menggunakan metode fuzzy sebagai pengelompokan pada nilai indeks prestasi Mahasiswa menjadi tiga bagian yaitu memuaskan, sangat memuaskan, dan cumlaude dan penelitian ini dapat memberikan kontribusi keilmuan pada penelitian K-Nearest Neighbor bagi peneliti lain untuk diterapkan pada kasus penelitian lain. Dan membantu pihak universitas untuk memberikan peringatan dini dan konsultasi lanjut bagi Mahasiswa yang kemungkinan tepat waktu dan tidak tepat waktu.

2. METODE PENELITIAN

Tahapan penelitian merupakan kegiatan penelitian yang dilakukan secara terencana, teratur, dan sistematis untuk mencapai tujuan tertentu. Tahapan penelitian ini juga merupakan pengembangan dari kerangka penelitian, dan terbagi lagi menjadi beberapa sub menu bagian. Gambar 1 berikut merupakan tahapan penelitian yang dilakukan:



Gambar 1 Tahapan Penelitian

Fuzzyfication Kriteria Indeks Prestasi

Indeks Prestasi Kumulatif (I)

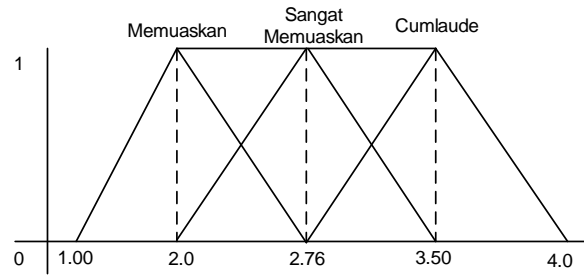
Nilai

2,76 – 4,00

I[1]=3,50

Cumlaude

2,00 - 3,50	I[1]=2,76	Sangat Memuaskan
1,00 -2,76	I[1]=2,00	Memuaskan



Gambar. 2 Grafik Kriteria IP [1]

Fungsi keanggotaan untuk himpunan MEMUASKAN:

$$\text{Memuaskan } [i] = \begin{cases} 0 & i \leq 1,00 \text{ atau } i \geq 2,76 \\ \frac{i-1}{2-1} & 1 \leq i \leq 2,00 \\ \frac{2,76-i}{2,76-2,00} & 2,00 \leq i \leq 2,76 \end{cases}$$

Fungsi keanggotaan untuk himpunan SANGAT MEMUASKAN:

$$\text{Sangat Memuaskan } [i] = \begin{cases} 0 & i \leq 2,0 \text{ atau } i \geq 3,5 \\ \frac{i-2,0}{2,76-2,00} & 2,0 \leq i \leq 2,76 \\ \frac{3,5-i}{3,5-2,76} & 2,76 \leq i \leq 3,5 \end{cases}$$

Fungsi keanggotaan untuk himpunan CUMLAUDE:

$$\text{Cumlaude } [i] = \begin{cases} 0 & i \leq 2,76 \text{ atau } i \geq 4,0 \\ \frac{i-2,76}{3,5-2,76} & 2,76 \leq i \leq 3,5 \\ 1 & i \geq 3,5 \end{cases}$$

3. HASIL DAN PEMBAHASAN

Tahap penjelasan hasil penelitian ini dilakukan setelah preprocess data telah selesai di ubah, kemudian dari data yang telah di ubah dalam bentuk fuzzy dilakukan proses pengelompokan terhadap atribut Indeks Prestasi Semester Satu, Dua, Tiga, Empat, dan Asal SMA. Pengelempokan ini dimaksudkan agar dapat mengurangi kesalahan-kesalahan yang timbul pada saat mengoperasikan sistem. Sistem yang dibuat ini digunakan dapat lebih baik karena belum adanya aplikasi prediksi tentang tingkat kelulusan Mahasiswa. Berikut akan diberikan penjelasan hasil penelitian yang ditulis secara rinci.

3.1. Pengolahan Data Awal

Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 1.685 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (*preparation data*). Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan adalah sebagai berikut:

1. *Data validation*, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*). *Missing data* terlihat

2. *Data integration and Transformation*, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data di transformasikan ke dalam *software RapidMiner*.
3. *Data size reduction and dicrtization*, untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informatif. Dalam penelitian ini atribut yang tidak relevan seperti indeks prestasi semester satu, dua, tiga dan empat diubah menjadi bentuk *fuzzyfication* dan SMA dikelompokkan menjadi angka 0-3.

3.2. Prediksi K-Nearest Neighbor

Pada Tahap Ini Penulis Menjelaskan tentang prediksi K-Nearest Neighbor yang dilakukan secara manual/ menggunakan excel, terdapat 50 data sampel angkatan 2008 yang merupakan data training sedangkan 1 data sampel Mahasiswa angkatan 2013 yang merupakan data testing. Menentukan terlebih dahulu K/ Jumlah tetangga terdekat dari prediksi kelulusan Mahasiswa, k yang digunakan dalam penelitian ini adalah 10. Menurut Hastie et al (2008), dengan k=5 atau 10 dapat digunakan untuk memperkirakan tingkat kesalahan yang terjadi, sebab data training pada setiap fold cukup berbeda dengan data training yang asli. Secara keseluruhan, 5 atau 10- fold cross validation sama-sama direkomendasikan dan disepakati bersama [4].

Tabel. 1 Data Testing Sebelum Preprocess

NPM	Nama	IP S-1	IP S-2	IP S-3	IP S-4	SPMB	Asal SMA	JK	Masa Studi
13312055	Budi Andriawan	3,6	3,38	3,2	3,05	330	SMK N 4 BDL	Laki-Laki	4 thn 3 bln 20 hr

Tabel. 2 Data Testing Sesudah Preprocess

NPM	Nama	IP S-1	IP S-2	IP S-3	IP S-4	SPMB	Asal SMA	JK	Masa Studi
13312055	Budi Andriawan	1	0,837	0,594	0,391	330	1	Laki-Laki	4 thn 3 bln 20 hr

Setelah Melakukan Proses *Preprocess data* pada *data training* dan *data testing*, langkah selanjutnya yang akan dilakukan yaitu menghitung jarak/ *distance* pada data sampel. Dalam penelitian ini menggunakan perhitungan jarak *euclidean distance* seperti pada persamaan 1 berikut:

$$D(x,y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad (1)$$

Hasil Perhitungan Jarak *Eulclidean Distance* Sampel untuk Data D4 dan Data D51 yang merupakan Data Testing dapat dilihat pada tabel 3 berikut:

Tabel. 3 Data Sampel Dengan Jarak *Euclidean*

No.	Nama	Distance
1	D04,D51	0,36
2	D05,D51	0,05
3	D13,D51	0,70
4	D14,D51	0,45
5	D33,D51	0,65
6	D40,D51	0,23
7	D32,D51	0,75
8	D16,D51	1,14
9	D46,D51	0,95
10	D49,D51	0,45

Dari 10 jarak yang telah didapatkan dari perhitungan menggunakan *Euclidean distance*, selanjutnya adalah menentukan prediksi menggunakan kategori mayoritas yang didapat dari klasifikasi data training ditunjukkan pada tabel 4 berikut:

Tabel. 4 Penentuan Prediksi

No.	Nama	Distance	Ranking	Klasifikasi
1	D16,D51	1,14	1	Terlambat
2	D46,D51	0,95	2	Tepat Waktu

No.	Nama	Distance	Ranking	Klasifikasi
3	D32,D51	0,75	3	Tepat Waktu
4	D13,D51	0,7	4	Tepat Waktu
5	D33,D51	0,65	5	Tepat Waktu
6	D14,D51	0,45	6	Tepat Waktu
7	D49,D51	0,45	7	Terlambat
8	D4,D51	0,36	8	Tepat Waktu
9	D40,D51	0,23	9	Terlambat
10	D5,D51	0,05	10	Tepat Waktu

Data Training yang sudah dilakukan perhitungan K-NN dengan Data Testing didapatkan kategori Lulus akan Tepat Waktu berjumlah 7, Sedangkan kategori Lulus akan Terlambat berjumlah 3. Data yang digunakan untuk proses perhitungan Fuzzy dan KNN merupakan Data yang telah diproses secara *fuzzyfication*. Tabel 5 berikut merupakan data indeks prestasi yang digunakan untuk proses prediksi kelulusan Mahasiswa.

Tabel. 6 Kelompok Data IP Semester 1

No	Nama	IP S-1	Fungsi Keanggotaan		
			Memuaskan	Sangat Memuaskan	Cumlaude
1	Adik Abdul Rohim	2,81	0,000	0,932	0,067
2	Agus Ade Putra	3,24	0,000	0,351	0,648
3	Ana Francisca	3,33	0,000	0,229	0,770
4	Apriyani	3,71	0,000	0,000	1,000
5	Bayu Adi Kurnia	2,33	0,565	0,434	0,000
6	Desi Jumiarti	2,38	0,5	0,5	0,000
7	Dwi Anita	2,86	0,000	0,135	0,864
8	Edwin Novanda	2,14	0,815	0,184	0,000
9	Erick Setiawan	2,86	0,000	0,135	0,864
10	Febi Herumanika	2,05	0,934	0,065	0,000

Pengujian model dalam penelitian ini menggunakan *Cross Validation* adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi. Dengan menggunakan *cross validation* akan dilakukan percobaan sebanyak k . Data yang digunakan dalam percobaan ini adalah data *training* untuk mencari nilai *error rate* secara keseluruhan. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau *10-fold Cross Validation*.

4. KESIMPULAN

Hasil penelitian terhadap Penerapan Metode Fuzzy dan K-Nearest Neighbor untuk prediksi kelulusan Mahasiswa maka dapat disimpulkan: Telah berhasil melakukan pemodelan untuk prediksi kelulusan Mahasiswa menggunakan algoritma Fuzzy dan K-Nearest Neighbor sebagai acuan Mahasiswa dalam meningkatkan mutu dan kualitas Mahasiswa kedepannya. Percobaan ini memiliki akurasi prediksi sebesar 77,35% dengan *true positive* berjumlah 1138 data dan *true negative* berjumlah 163 data. *class precision* untuk prediksi lulus tepat waktu sebesar 79,86% serta untuk prediksi terlambat Sebesar 63,42%, memiliki *class recall* untuk *true* tepat sebesar 92,37% dan *true* Terlambat sebesar 36,22%.

UCAPAN TERIMA KASIH

Kami ucapkan terimakasih kepada Universitas Teknokrat Indonesia yang telah menyediakan sumber data sehingga dapat digunakan untuk proses penelitian ini.

REFERENSI

- [1] Wantoro A, Priandika AT, "Komparasi Perhitungan Pemilihan Mahasiswa Terbaik Menggunakan Metode Statistik Klasik Dengan Logika Fuzzy (Tsukamoto Dan Mamdani) Studi Kasus STMIK Teknokrat", *Sistem Informasi STMIK Teknokrat*, 2017, Vol. 15 No. 1.
- [2] Iin E, "Perbandingan algoritma C4.5 dan K-Nearest Neighbor untuk Prediksi Status Keaktifan Studi Mahasiswa", 2008.
- [3] Gorunescu, F, "Data Mining Concepts Models and Techniques", 2011, *Craiova: Springer*.

- [4] Trevor H, Robert T, Friedman J, “The Elements of Statistical Learning Data Mining, Inference, and Prediction”, 2008, California: Springer.
- [5] Ahmad I, Hermadi I, Arkeman Y, “Financial Feasibility Study Of Waste Cooking Oil Utilization for Biodiesel Production Using ANFIS”, *TELKOMNIKA*, 2015, Vol. 13 No 3.
- [6] Karamouiz, Vrettos, “Sensitiv Analysis of Neural Network for Identifying the Factors For College Students Success”, 2009.
- [7] Larose, Daniel. T, “Discovering Knowledge in Data”, 2005, New Jersey: John Willey & Sons, Inc.
- [8] Mustakim, Oktaviani GF, “Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa”, *Jurnal Sains Teknologi Industri*, 2016, Vol. 13, No. 2, pp: 195-202.
- [9] Ricky NI, Kusrini , Arief MR, “Analisis Prediksi Tingkat Pengunduran Diri Mahasiswa STIKOM UYELINDO dengan Metode K-Nearest Neighbor”, Kupang: Magister Teknik Informatika, 2014.
- [10] Qudri, Kalyankar, “Drop Out Feature of Student Data for Academic Performance Using Decision Tree techniques”, 2010.
- [11] Pandie, Emerensye SY, “Implementasi Algoritma Data Mining K-Nearest Neighbour (KNN) Dalam Pengambilan Keputusan Pengajuan Kredit”, *Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana:Kupang*, 2012.
- [12] Suhartina, Ernastuti, “Graduation Prediction of Gunadarma University Students Using Algorithm and Naïve Bayes C4.5 Algorithm”, 2010.
- [13] Wu X, “The Top Ten Algorithms in Data Mining”, 2009, New York: CRC Press.
- [14] Mohammad M, Amad I, Fernando Y, “Pemetaan Potensi Pariwisata Kabupaten Waykanan Menggunakan Algoritma Dijkstra”, *Jurnal Komputer Terapan*, 2017, Vol. 1 No. 2, pp: 169-178.

BIBLIOGRAFI PENULIS



Imam Ahmad, M.Kom lahir di Punggur, Lampung Tengah. Lulus S1 dari Program studi Sistem Informasi STMIK Abulyatama Banda Aceh, melanjutkan program pasca sarjana di Institut Pertanian Bogor dengan Jurusan Ilmu Komputer. Sekarang aktif mengajar pada Universitas Teknokrat Indonesia, Bandar Lampung.



Heni Sulistiyani S.Kom., M.Kom lahir di Polaman, 12 Oktober 1986. Menempuh pendidikan sarjana di STMIK Teknokrat, Jurusan Teknik dan melanjutkan pasca sarjana di Informatika Institut Teknologi Sepuluh Nopember Jurusan Sistem Informasi, aktif mengajar di Universitas Teknokrat Indonesia. Sekarang sedang menempuh program Doktor di Universitas Lampung.



Hendrik Saputra, S.Kom. lahir di Bandar Lampung pada 1 Oktober 1995, Lulus S1 dari Program Studi Informatika, Universitas Teknokrat Indonesia tahun 2018