

**ANALISIS HASIL UJIAN NASIONAL BERDASARKAN
KARAKTERISTIK SEKOLAH DENGAN
ALGORITMA KNNC4.5**

*ANALYSIS OF NATIONAL EXAMINATION RESULTS
BASED ON SCHOOL CHARACTERISTICS
WITH KNNC4.5 ALGORITHM*

DESSY SANTI



**PROGRAM PASCASARJANA
UNIVERSITAS HASANUDDIN
MAKASSAR
2013**

**ANALISIS HASIL UJIAN NASIONAL BERDASARKAN
KARAKTERISTIK SEKOLAH DENGAN
ALGORITMA KNNC4.5**

ANALYSIS OF NATIONAL EXAMINATION RESULTS
BASED ON SCHOOL CHARACTERISTICS
WITH KNNC4.5 ALGORITHM

DESSY SANTI

**PROGRAM PASCASARJANA
UNIVERSITAS HASANUDDIN
MAKASSAR**

2013

PERNYATAAN KEASLIAN TESIS

Yang bertanda tangan di bawah ini :

Nama : Dessy Santi
Nomor Pokok : P2700211458
Program Studi : Teknik Elektro
Konsentrasi : Teknik Informatika

Menyatakan dengan sebenarnya bahwa tesis yang saya tulis ini benar-benar merupakan hasil karya sendiri, bukan merupakan pengambilalihan tulisan atau pemikiran orang lain. Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan tesis ini hasil karya orang lain, saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, Juli 2013

Yang menyatakan,

Dessy Santi

KATA PENGANTAR

Puji Syukur kepada ALLAH Bapa yang Maha Kuasa, Putranya Yesus Kristus dan Roh Kudus serta Bunda Maria yang telah memberikan rahmatnya sehingga penulis dapat menyelesaikan Tesis dengan judul, **“ANALISIS HASIL UJIAN NASIONAL BERDASARKAN KARAKTERISTIK SEKOLAH DENGAN ALGORITMA KNN4.5 ”**.

Penulis menyadari banyaknya kendala dalam penyusunan tesis ini, namun demikian penulis dapat menyelesaikannya berkat bantuan dari berbagai pihak baik secara moril maupun materil. Karena itu dalam kesempatan ini perkenankan penulis menyampaikan ucapan terima kasih dan penghargaan setinggi - tingginya kepada :

1. Kedua orang tua tercinta, bapak Prof. Dr. Maxinus Jaeng, M.Pd, Mama Josina Bugid yang membesarkan dan mendidik penulis serta senantiasa mendoakan dengan tulus ikhlas dan senantiasa mendukung sehingga. Serta tak lupa kepada saudaraku Sylviana jaeng dan Theresia sestiawati Jaeng, Yanto dan ponakanku tercinta Aldo yang menjadi salah satu motivasi penulis dalam menyelesaikan tesis ini.
2. **Prof.Dr.Ir.H. Nadjamuddin Harun, MS**, sebagai Ketua Komisi Penasihat dan **Dr.Adnan,ST.,MT** sebagai anggota Komisi Penasihat atas bantuan dan bimbingan yang telah diberikan mulai dari proposal dan pelaksanaan penelitian hingga penulisan tesis

3. Tim Penguji **Dr .Ir. Zahir Zainuddin,M.Sc, Drs. Suarga, M.Sc.,M.Math., Ph.D**, dan **Dr. Armin Lawi, S.Si., M.Eng** yang telah memberikan saran, kritik dan pertimbangan dalam penyempurnaan Tesis ini.
4. **Prof.Dr.Ir.Salama Manjang,MT** selaku ketua Program Studi S2 Teknik Elektro Universitas Hasanuddin.
5. Kepada Nenek tercinta yang selalu perhatian serta medoakan penulis dan semua keluarga dan teman-teman serta kerabat di Makassar dan Palu yang penulis tidak dapat menyebutkan satu-persatu yang selalu membantu dan mendukung dalam masa pendidikan dan penyelesaian Tesis.
6. Kepada semua dosen-dosen Universitas Hasanuddin, yang telah memberikan Ilmunya kepada penulis tanpa pamrih, serta staf pegawai yang sudah banyak membantu secara adminitrasi hingga selesainya tugas akhir.
7. Teman-teman PascaMelekIT 2011 something special dalam pencarian tuntutan keilmuan di tengah kebersamaan yang tak terhingga dan tak terlupakan.

Makassar, Juli 2013

Dessy Santi

Abstrak

Dessy Santi, Analisis Hasil Ujian Nasional Berdasarkan Karakteristik Sekolah dengan Algoritma KNNC4.5, dibimbing oleh : **Nadjamuddin Harun** dan **Adnan**

Kebijakan pemerintah untuk memberikan bantuan kepada sekolah sangat berpengaruh terhadap tingkat kelulusan hasil ujian nasional setiap sekolah maka pemerintah perlu memberikan bantuan dengan tepat bagi SMA Negeri/Swasta yang membutuhkan perbaikan karakteristik sekolah.

Penelitian ini bertujuan (1) mengetahui gambaran klasifikasi tingkat kelulusan berdasarkan karakteristik sekolah (2) Menguji akurasi algoritma KNNC4.5 terhadap banyak data tes dan data training. Penelitian ini merupakan penelitian historis yang bersifat aplikatif sehingga dilakukan dengan metode studi pustaka, metode pengumpulan data dan pembuatan aplikasi berdasarkan hasil analisa dari metode KNNC4.5 dengan ,membagi atribut berdasarkan karakteristik sekolah yaitu data yang bernilai kuantitatif diproses dengan algoritma KNN dan data kualitatif diproses dengan algoritma C4.5 kemudian hasil kedua algoritma dikombinasi menjadi algoritma KNNC4.

Penelitian ini menghasilkan Kombinasi algoritma KNN dan algoritma C4.5 menjadi algoritma KNNC4.5 yang menghasilkan prediksi tingkat kelulusan dengan kelas 100% dan <100% dan factor-faktor apa saja yang mempengaruhi tingkat kelulusan yang mempunyai nilai akurasi 81% dengan inputan nilai k untuk proses KNN untuk nilai k yang semakin besar maka akurasi akan menurun begitu pula sebaliknya. Aturan kelulusan berdasarkan pohon keputusan dari proses C4.5. Pada akhirnya hasil penelitian ini mampu memberikan kontribusi pada pemerintah untuk kesuksesan ujian nasional sehingga meningkatnya mutu dan, kualitas pendidikan.

Kata Kunci : karakteristik, klasifikasi Nearest Neighbor (KNN), algoritma C4.5

Abstract

Dessy Santi, Analysis Of The Results Of National Examinations Algorithm Based On Characteristics Of School With KnnC4.5, Author by : Nadjamuddin Harun and Adnan

The government's policy for giving out aid to schools is influential on the national graduation rate of each school examination results, the government needs to provide the appropriate assistance to high schools / private schools in need of improvement characteristics.

This study aims to (1) determine the classification description based karakteristik school graduation rates (2) Test the accuracy of the algorithm KNNC4.5 many test data and training data. This study is a historical research that is applicable to do with library research methods, data collection methods and the creation of applications based on the analysis of the KNNC4.5 method, dividing the attributes based on characteristics of the school that is valuable quantitative data calculated by the KNN algorithm and qualitative data calculated by C4.5 algorithms then combine the results of these two algorithms into algorithms KNNC4.

The Study will generate Combine of KNN algorithm and C4.5 algorithm to become KNNC4.5 algorithm and generate predictions graduation rate of 100% and <100%, and the factors that influence graduation rates that have value 81% accuracy with input values of k greater the accuracy will not get better and graduation rule based decision tree of the C4 .5. eventually be able to contribute to the success of the government for the national exam and thus increase the quality, the quality of education.

Keywords: Characteristics, Nearest Neighbor Classification (KNN), C4.5 Algorithm

DAFTAR ISI

	halaman
HALAMAN JUDUL	i
HALAMAN PENGESAHAN.....	iii
HALAMAN PERNYATAAN.....	iv
KATA PENGANTAR.....	v
ABSTRAK	vii
ABSTRACT	viii
DAFTAR ISI	ix
DAFTAR TABEL	xii
DAFTAR GAMBAR	xiv
Bab I PENDAHULUAN	
A. Latar Belakang Masalah.....	1
B. Rumusan Masalah	3
C. Tujuan Penelitian.....	3
D. Manfaat Penelitian.....	4
E. Batasan Masalah	4
BAB II. TINJAUAN PUSTAKA	
A. Ujian Nasional SMA	5
B. Karakteristik Sekolah	6
C. Konsep Data Mining	8

D. Algoritma	9
E. Data	10
F. Klasifikasi	11
G. Klasifikasi Nearest Neighbor (KNN)	12
1. Proses K-Nearest Neighbor	13
2. Proses Modified KNN	14
H. Algoritma C4.5	17
I. Perhitungan Akurasi.....	20
J. Penelitian Yang Serupa	20
K. Kerangka Pikir	23

BAB III. METODE PENELITIAN

A. Waktu dan Lokasi Penelitian	24
B. Jenis Penelitian	24
C. Alat dan Bahan.....	24
D. Tahapan Rancangan Penelitian	25
1. Analisis Kebutuhan Sistem	26
2. Perancangan Sistem	26
3. Deskripsi Data	33
4. Perancangan Proses	35
E. Metode Pengujian Sistem	46

BAB IV. HASIL PENELITIAN DAN PEMBAHASAN

A. Gambaran Umum Sistem.....	47
B. Simulasi Algoritma	49
1. Algoritma K-Nearest Neighbor	49
2. Algoritma C4.5	57
3. Algoritma KNNC4.5	67
C. Implementasi Sistem	70
D. Pengujian Sistem	75

BAB V. PENUTUP

A. Kesimpulan	85
B. Saran.....	86

Daftar Pustaka

Lampiran

DAFTAR GAMBAR

	halaman
GAMBAR 2.1 PROSES KNOWLEDGE DISCOVERY DATA	10
GAMBAR 3.1 TAHAPAN RANCANGAN PENELITIAN	25
GAMBAR 3.2. USE CASE DIAGRAM.....	27
GAMBAR 3.3. ACTIVITY DIAGRAM INPUT DATA SEKOLAH	29
GAMBAR 3.4. ACTIVITY DIAGRAM PROSES ALGORITMA KNN.....	30
GAMBAR 3.5. ACTIVITY DIAGRAM PROSES C4.5.....	31
GAMBAR 3.6. CLASS DIAGRAM	32
GAMBAR 3.7. FLOWCHART SISTEM KESELURUHAN	36
GAMBAR 3.8. FLOWCHART PROSES KLASIFIKASI DATA	38
GAMBAR 3.9. FLOWCHART MENGHITUNG VALIDITAS	40
GAMBAR 3.10. FLOWCHART MENGHITUNG EUCLIDEAN.....	42
GAMBAR 3.11 FLOWCHART MENGHITUNG WEIGH VOTING	43
GAMBAR 3.12. FLOWCHART KLASIFIKASI ALGORITMA C4.5	45
GAMBAR 4.1. GAMBARAN UMUM SISTEM.....	47
GAMBAR 4.2. PROSES KLASIFIKASI DATA.....	48
GAMBAR 4.3. P0HON KEPUTUSAN PERHITUNGAN LEVEL 0.....	60
GAMBAR 4.4. P0HON KEPUTUSAN PERHITUNGAN LEVEL 1.....	62
GAMBAR 4.5. P0HON KEPUTUSAN PERHITUNGAN LEVEL 1.....	65
GAMBAR 4.6 HALAMAN UTAMA.....	70
GAMBAR 4.7. FORM INPUT DATA SEKOLAH	71

GAMBAR 4.8. FORM INPUT DATA TESTING	72
GAMBAR 4.9. FORM PROSES KNN	72
GAMBAR 4.10. FORM PROSES C4.5	73
GAMBAR 4.11. FORM PROSES KNN C4.5	74
GAMBAR 4.12. FORM OUTPUT TINGKAT KELULUSAN	74
GAMBAR 4.13. GRAFIK TINGKAT AKURASI	82

DAFTAR TABEL

	halaman
TABEL 4.1 TABEL DATA TESTING DAN DATA TRAINING.....	50
TABEL 4.2 TABEL CARA HITUNG VALIDITAS.....	51
TABEL 4.3. TABEL HASIL PERHITUNGAN VALIDITAS	53
TABEL 4.4. TABEL HASIL PERHITUNGAN EUCLIDEAN	54
TABEL 4.5. TABEL HASIL PERHITUNGAN WEIGHT VOTING.....	55
TABEL 4.6. TABEL PENETUAN KELAS	56
TABEL 4.7. TABEL KASUS PENENTUAN KELULUSAN	57
TABEL 4.8. TABEL PERHITUNGAN LEVEL 0	59
TABEL 4.9. TABEL PERHITUNGAN LEVEL 1	61
TABEL 4.10. TABEL PERHITUNGAN LEVEL 2.....	63
TABEL 4.11 TABEL OUTPUT KOMBINASI ALGORITMA KNNC4.5	68
TABEL 4.12 TABEL PREDIKSI PADA SMA NEGERI 2 PALU.....	69
TABEL 4.13. TABEL HASIL PENGUJIAN BALCK BOX.....	75
TABEL 4.14. TABEL HASIL PENGUJIAN TEHADAP NILAI K.....	80
TABEL 4.15. TABEL PENGUJIAN AKURASI	83

BAB I

PENDAHULUAN

A. Latar Belakang

Hasil ujian Nasional sebagai peta dan pintu masuk perbaikan kualitas pendidikan melalui intervensi kebijakan ditingkat satuan pendidikan. Tidak hanya mata pelajaran yang diujikan. Peta = f (kewilayahan , mata pelajaran, sumber daya pendidikan). Menurut peraturan pemerintah No. 19 tahun 2005. Ps. 68, hasil ujian negara digunakan sebagai salah satu pertimbangan untuk :

1. Pemetaan mutu program dan atau satuan pendidikan.
2. Dasar seleksi masuk jenjang pendidikan berikutnya.
3. Penentuan kelulusan peserta didik dari program/ satuan pendidikan
4. Pembinaan dan pemberi bantuan kepada satuan pendidikan dalam upaya untuk meningkatkan mutu pendidikan[1].

Pemerintah sebagai pengambil kebijakan, harus arif bijaksana serta memberikan rasa aman kepada para peserta didik, bahwa mereka bisa lulus Ujian Nasional. Hal ini bisa diwujudkan, dengan membuat semua sekolah di negeri ini menjadi layak pakai dan memenuhi standar sebagai sebuah sekolah. Sangat tidak mungkin bila pemerintah mendapatkan hasil

yang baik dalam Ujian Nasional, bila pemerintah sendiri belum memberikan kualitas pendidikan yang layak. Hal ini mulai dari sarana prasarana hingga pemerataan guru yang berkualitas. Selama ini, selalu terjadi kesenjangan. Masih banyak sekolah yang belum bisa disebut sebagai sebuah sekolah. Ini dibuktikan dengan aneka temuan, bahwa banyak sekolah yang tidak memenuhi standar. Data Kementerian Pendidikan dan Kebudayaan (Kemendikbud) menyebutkan, ada sekitar 161 ribu sekolah rusak, 45% dari gedung sekolah rusak tersebut mengalami rusak berat. Sekolah-sekolah seperti ini, pada umumnya berada di daerah-daerah terpencil.

Oleh karena kebijakan pemerintah untuk memberikan bantuan kepada sekolah sangat berpengaruh terhadap tingkat kelulusan maka pemerintah harus mempunyai acuan dalam memberikan intervensi kebijakan tersebut. Salah satu cara pemerintah dalam mengambil keputusan adalah melakukan analisis hasil ujian nasional berdasarkan karakteristik sekolah dengan menerapkan konsep penalaran berbasis kasus (case based reasoning). Konsep ini pada dasarnya adalah membandingkan kasus yang terjadi dengan database kasus yang ada, untuk mencari solusi kasus baru dengan menirukan solusi yang diambil dari kasus sebelumnya. Model klasifikasi yang digunakan adalah dengan menggunakan metode algoritma *Klasifikasi-Nearest Neighbor* (K-NN) dan Algoritma C4.5 untuk memperoleh akurasi yang tepat.

Dengan demikian diharapkan hasil penelitian ini bisa menjadi salah satu bahan acuan pemerintah dalam pengambilan keputusan untuk memberikan kebijakan dan bantuan pada sekolah-sekolah yang tingkat kelulusannya rendah, dalam rangka memperbaiki mutu dan kualitas pendidikan yang berpengaruh pada tingkat kelulusan.

B. Rumusan Masalah

Bagaimana mengetahui gambaran klasifikasi tingkat kelulusan setiap SMU Negeri/Swasta berdasarkan karakteristik Sekolah.

C. Tujuan Penelitian

1. Untuk mengetahui gambaran klasifikasi tingkat kelulusan setiap SMU Negeri/Swasta berdasarkan karakteristik Sekolah sehingga adanya perbaikan pada karakteristik sekolah.
2. Pengujian akurasi algoritma KNN C4.5 terhadap banyak data tes dan hasil prediksi tingkat kelulusan yang akurat dengan komputasi yang cepat.

D. Manfaat Penelitian

1. Dapat membantu pemerintah menentukan sekolah yang tepat untuk memperoleh bantuan dan intervensi kebijakan.
2. Meningkatkan mutu dan perbaikan kualitas pendidikan.
3. Metode Algoritma KNN C4.5 dapat digunakan dalam menganalisa dan memprediksi kasus yang berbeda seperti kesehatan dan pendidikan.

E. Batasan Masalah

Agar dapat mencapai sasaran dan tujuan yang diharapkan maka permasalahan dibatasi pada :

1. Prediksi tingkat kelulusan hanya berdasarkan karakteristik sekolah
2. Karakteristik sekolah yang dimaksud adalah seluruh SMA swasta maupun negeri pada propinsi Sulawesi Tengah.
3. Analisis pola klasifikasi sekolah menggunakan Algoritma KNN untuk mengetahui tingkat kelulusan di masing-masing SMA di Sulawesi Tengah.
4. Hasil Klasifikasi tingkat kelulusan berdasarkan karakteristik sekolah akan menghasilkan prediksi tingkat kelulusan pada setiap sekolah ditahun berikutnya menggunakan algoritma c4.5.

BAB II

TINJAUAN PUSTAKA

A. Ujian Nasional SMA

Ujian Nasional adalah sistem evaluasi standar pendidikan dasar dan menengah secara nasional dan persamaan mutu tingkat pendidikan antar daerah yang dilakukan oleh Pusat Penilaian Pendidikan, Depdiknas di Indonesia berdasarkan *Undang-Undang Republik Indonesia No.20 Tahun 2003* menyatakan bahwa dalam rangka pengendalian mutu pendidikan secara nasional dilakukan evaluasi sebagai bentuk akuntabilitas penyelenggara pendidikan kepada pihak-pihak yang berkepentingan. Lebih lanjut dinyatakan bahwa evaluasi dilakukan oleh lembaga yang mandiri secara berkala, menyeluruh, transparan, dan sistematis untuk menilai pencapaian standar nasional pendidikan dan proses pemantauan evaluasi tersebut harus dilakukan secara berkesinambungan.[1]

Proses pemantauan evaluasi tersebut dilakukan secara terus menerus dan berkesinambungan pada akhirnya akan dapat membenahi mutu pendidikan. Pembenahan mutu pendidikan dimulai dengan penentuan standar.

Penentuan standar yang terus meningkat diharapkan akan mendorong

peningkatan mutu pendidikan, yang dimaksud dengan penentuan standar pendidikan adalah penentuan nilai batas (*cut off score*). Seseorang dikatakan sudah lulus/kompeten bila telah melewati nilai batas tersebut berupa nilai batas antara peserta didik yang sudah menguasai kompetensi tertentu dengan peserta didik yang belum menguasai kompetensi tertentu. Bila itu terjadi pada ujian nasional atau sekolah maka nilai batas berfungsi untuk memisahkan antara peserta didik yang lulus dan tidak lulus disebut batas kelulusan, kegiatan penentuan batas kelulusan disebut *standard setting*.

B. Karakteristik Sekolah

karakteristik diambil dari bahasa Inggris yakni *characteristic*, yang artinya mengandung sifat khas. Ia mengungkapkan sifat-sifat yang khas dari sesuatu Dalam kamus lengkap psikologi karya Chaplin, dijelaskan bahwa karakteristik merupakan sinonim dari kata karakter, watak, dan sifat yang memiliki pengertian diantaranya :

1. Intergrasi atau sintese dari sifat-sifat individual dalam bentuk suatu untas atau kesatuan
2. Suatu kualitas dan sifat yang tetap terus-menerus dan kekal yang dapat dijadikan cirri untuk mengidentifikasikan seorang pribadi, suatu objek, suatu kejadian.

3. Kepribadian seseorang, dipertimbangkan dari titik pandangan etis atau moral.

Jadi di antara pengertian-pengertian di atas sebagaimana yang telah dikemukakan oleh Chaplin, dapat disimpulkan bahwa karakteristik itu adalah suatu sifat yang khas, yang melekat pada seseorang atau suatu objek, dalam hal ini adalah sekolah menengah umum.

Sekolah Menengah Atas dalam pendidikan formal di Indonesia, merupakan jenjang pendidikan menengah setelah menamatkan Sekolah Menengah Pertama (SMP) atau yang sederajat. Sekolah Menengah Atas diselesaikan dalam kurun waktu 3 tahun, yaitu mulai kelas 10 sampai kelas 12. Pada tahun kedua (di kelas 11), siswa Sekolah Menengah Atas, wajib memilih jurusan yang ada, yaitu Sains, Sosial, atau Bahasa. Pada akhir tahun ketiga (di kelas 12), siswa diwajibkan mengikuti Ujian Nasional yang mempengaruhi kelulusan atau tidaknya siswa. Setelah lulus (tamat)

Sekolah Menengah Atas dapat melanjutkan pendidikan ke perguruan tinggi. Umumnya pelajar Sekolah Menengah Atas berusia 16-18 tahun.

Sekolah Menengah Atas tidak termasuk program wajib belajar pemerintah seperti SD 6 tahun serta SMP 3 tahun. Mulai tahun 2005, di beberapa daerah di Indonesia, Sekolah Menengah Atas telah diikutkan sebagai program wajib belajar 12 tahun yang diselenggarakan oleh pemerintah maupun swasta. Pengelolaan Sekolah Menengah Atas negeri di Indonesia yang sebelumnya berada di bawah Departemen Pendidikan Nasional, setelah diberlakukannya otonomi daerah pada tahun 2001, kini menjadi

tanggung jawab pemerintah daerah kabupaten/kota. Sedangkan Departemen Pendidikan Nasional hanya berperan sebagai regulator dalam bidang standar nasional pendidikan. Jadi Karakteristik sekolah khususnya Menengah Umum adalah suatu sifat yang khas suatu dan sifat yang melekat yang dapat dijadikan ciri untuk mengidentifikasi sesuatu yang terdapat dalam sekolah, status, *jumlah siswa, jumlah rombel, jumlah guru, Laboratorium, ruang praktek, perpustakaan.*

C. Konsep Data Mining

Data Mining merupakan suatu disiplin ilmu baru dalam teknologi komputasi dan informatika yang didefinisikan sebagai metode untuk mengekstrak pengetahuan yang sifatnya implisit dan sebelumnya tidak diketahui dari data yang berskala besar. Peran utama dalam data mining adalah estimasi dan prediksi menggunakan algoritma Linear Regression, Neural Network, support Vector Machine; klasifikasi menggunakan algoritma Naïve Bayes, k-Nearest Neighbor, C4.5 CART, Linear Discriminant Analyst; klasterisasi menggunakan algoritma K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means; dan asosiasi menggunakan algoritma FP-Growth, A Priori dan Hase Based Assosiation. (Kursini, 2009)[2].

Metode yang digunakan dalam data mining berupa metode pembelajaran (*supervised learning*) dan metode tanpa pembelajaran (*UnSupervised learning*). Metode pembelajaran meliputi peran estimasi, prediksi, klasifikasi dan asosiasi sedangkan metode tanpa pembelajaran meliputi klasterisasi. (Budi Santosa, 2007)[3].

D. Algoritma

Definisi Algoritma adalah langkah-langkah logis penyelesaian masalah yang disusun secara sistematis dan logis. Penemu kata Algoritma sendiri adalah nama Abu Ja'far Mohammed Ibnu Musa Al Khowarizmi, ilmuan Persia yang menulis kitab al jabr w'almuqabala sekitar tahun 825 M.

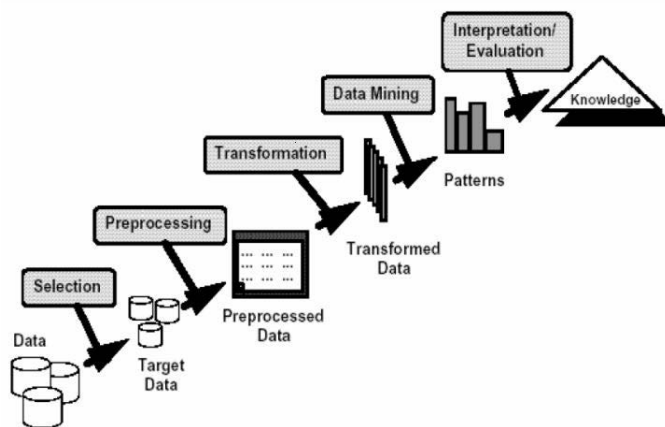
Menurut Donald E. Knuth dalam bukunya yang berjudul “ The Art of Computer Programming”[4], algoritma memiliki 5 ciri-ciri penting yakni :

1. Algoritma harus berhenti setelah melakukan sejumlah langkah terbatas.
2. Setiap langkah algoritma harus didefinisikan dengan tepat dan tidak bermakna ganda (ambiguous).
3. Algoritma memiliki nol atau lebih masukan (input)
4. Algoritma memiliki satu atau beberapa keluaran (output).
5. Algoritma harus efektif.

E. Data

Data sering disebut sebagai bahan mentah informasi. Tapi menurut Murdick, dkk (1984) merumuskan bahwa data adalah fakta yang tidak sedang digunakan pada proses keputusan, biasanya dicatat dan diarsipkan tanpa maksud untuk segera diambil kembali untuk pengambilan keputusan.

Knowledge discovery data (KDD) adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang ditemukan bersifat sah, baru dapat bermanfaat dan dapat dimengerti (Usama Fayyad, 1996). Gambar 1 menunjukkan proses dari KDD.



Gambar 1. Proses *Knowledge discovery in database (KDD)*

(Usama Fayyad, 1996)

F. Klasifikasi

Klasifikasi merupakan salah satu teknik data mining yang memiliki kemampuan untuk melakukan proses klasifikasi data. Klasifikasi bisa digunakan untuk menemukan model atau fungsi yang membedakan kelas data. Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang labelnya tidak diketahui.

Klasifikasi merupakan suatu teknik dengan melihat kelakuan atribut dari kelompok yang telah didefinisikan, Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan-aturan tersebut digunakan pada data-data baru yang diklasifikasi. Teknik ini menggunakan *supervised induction*, yang memanfaatkan kumpulan pengujian dari record yang terklasifikasi untuk menentukan kelas-kelas tambahan(Kusnawi, 2007)[5].

Klasifikasi merupakan proses untuk menyatakan suatu objek ke dalam salah satu kategori yang sudah didefinisikan sebelumnya. Tujuan dari klasifikasi ini adalah *record-record* yang sebelumnya belum termasuk dalam kategori dapat dinyatakan kelasnya secara akurat. Tahapan-tahapan klasifikasi terdiri dari :

1. Pembangunan model

Dalam tahapan ini dibuat model untuk menyelesaikan masalah klasifikasi data, model ini dibangun berdasarkan *training set*.

2. Penerapan model

Pada tahapan ini model yang sudah dibangun sebelumnya digunakan untuk menentukan *atribut* atau *class* dari sebuah data yang *atribut* atau *class*nya belum diketahui.

3. Evaluasi

Dalam tahapan ini hasil dari tahapan sebelumnya dievaluasi menggunakan parameter terukur untuk menentukan apakah model tersebut dapat diterima.

Pada proses klasifikasi terdapat beberapa metode, antara lain *decision tree*, *Bayesian*, *fuzzy*, *neural network*, *support vector machine (SVM)* dan *k-nearest neighbor* (Pramudiono, 2003)[6].

G. K-Nearest Neighbor (KNN)

Algoritma *K-Nearest Neighbor (KNN)* adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut.

Data pembelajaran diproyeksikan ke ruang dimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Sebuah titik pada ruang ini ditandai kelas *c* jika kelas *c* merupakan klasifikasi yang paling banyak ditemui pada *k* buah tetangga terdekat titik tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Euclidian[2]

1. Proses *K-Nearest Neighbor* (K-NN)

Algoritma *K-Nearest Neighbor*

1. Tentukan K (misalnya k = 3)
2. Hitung jarak antara data baru ke setiap label data
3. Tentukan k labeled data yang mempunyai jarak yang paling minimal
4. Klasifikasikan data baru ke dalam data yang mayoritas.

Menurut Agusta, 2007 bahwa prinsip kerja *K-Nearest Neighbor(KNN)* adalah mencari jarak terdekat antara data yang dievaluasi dengan k tetangga terdekatnya dalam data pelatihan. Persamaan perhitungan untuk mencari euclidian dengan d adalah jarak p adalah dimensi data dengan persamaan[7].

$$de = \sqrt{\sum_{1-i}^p (x_{2i} - x_{1i})^2} \dots \quad (2.1)$$

Dimana:

X1 : sample data uji

X2 : data uji

De : jarak

P : dimensi data

2. Proses Modified K-Nearest Neighbor

Modified K-Nearest Neighbor adalah menempatkan label kelas data sesuai dengan k divalidasi poin data yang sudah ditetapkan dengan perhitungan K-Nearest Neighbor (KNN) tertimbang (Hamid Parvin, 2008)[8], berikut proses dari Modified KNN

a. *Validitas Data Training*

Dalam algoritma MKNN, setiap data pada *data training* harus divalidasi terlebih dahulu pada awalnya. *Validitas* setiap data tergantung pada setiap tetangganya. Proses validasi dilakukan untuk semua data pada *data training*. Setelah dihitung *validitas* tiap data maka nilai *validitas* tersebut digunakan sebagai informasi lebih mengenai data tersebut.

Untuk menghitung *validitas* dari data pada *data training* tetangga terdekatnya perlu dipertimbangkan. Diantara tetangga terdekat data, *validitas* digunakan untuk menghitung jumlah titik dengan label yang sama untuk data tersebut. Persamaan yang digunakan dari setiap titik pada *data training* adalah seperti pada persamaan berikut :

$$Validitas(x) = \frac{1}{k} \sum_{l=1}^k S(lbl(x), (lbl(N_i(x)))) \dots \quad (2.2)$$

Dimana :

- k : Jumlah titik terdekat
- LBL(x) : kelas x
- N_i(X) : label kelas titik terdekat x

Fungsi S digunakan untuk menghitung kesamaan antara titik x dan data ke-i dari tetangga terdekat yang dituliskan dalam persamaan 2.3 didefinisikan fungsi ini.

$$S(a,b) = \begin{cases} 1 & a=b \\ 0 & a \neq b \end{cases} \dots (2.3)$$

Keterangan : a = kelas a pada data training

b = kelas lain selain a pada data training

b. *Weight voting KNN*

Weight voting KNN adalah salah satu variasi metode KNN yang menggunakan K tetangga terdekat, terlepas dari kelas data, tapi menggunakan *Weight voting* dari masing-masing data pada *data training*. Masing-masing data diberikan *Weight voting* yang biasanya sama dengan beberapa penurunan fungsi jarak dari data yang tidak diketahui. Sebagai contoh \, voting diatur sama dengan $1/(de + 1)$, dimana d e adalah jarak Eucledian. *Weight voting* ini kemudian dijumlahkan untuk setiap kelasnya, dan kelas dengan jumlah terbesar suara yang dipilih.

Dalam metode MKNN, pertama weight masing-masing tetangga dihitung dengan menggunakan $1/(de + 0,5)$. Kemudian, *validitas* dari tiap data training dikalikan dengan weight berdasarkan pada jarak Euclidian. Dalam metode MKNN, *Weight voting* tiap tetangga seperti persamaan 2.4.

$$W(i) = Validitas(i) \times \frac{1}{de + 0.5} \quad \dots (2.4)$$

Dimana:

$W(i)$: Perhitungan *Weight voting*

$Validitas(i)$: Nilai *Validitas*

De : jarak Euclidian

Teknik *Weight voting* ini mempunyai pengaruh yang lebih penting terhadap data yang mempunyai nilai *validitas* lebih tinggi dan paling dekat dengan data. Selain itu, dengan mengalikan *validitas* dengan jarak dapat mengatasi kelemahan dari setiap data yang mempunyai jarak dengan weight yang memiliki banyak masalah *outlier*.

Beberapa keuntungan dari metode K-Nearest Neighbor (KNN) adalah sebagai berikut:

- a. Sederhana dalam penggunaannya.
- b. Dapat menangani *data training* yang mengandung *noise*
- c. Efektif jika *data training* besar.

3. Pseudocode Modified K- Nearest Neighbor (KNN)

```

Output_label := MKNN (train_set, test_sample)
Begin
  For i := 1 to train_size
    Validitas(i) := Hitung validitas sampel ke-i;
  End for;
  Output_label := Weighted_KNN(Validitas, test_sample);
  Return Output_label;
End

```

H. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variable input dengan sebuah variable target.

Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi rule, dan menyederhanakan *rule* (Basuki & Syarif, 2003).

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut, (Kusirini, 2009) [2]:

1. Pilih atribut sebagai akar
2. Buat cabang untuk tiap-tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *Gain* digunakan rumus seperti tertera dalam persamaan 2.5.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \quad \dots(2.5)$$

Keterangan :

S : himpunan kasus

A : Atribut

N : jumlah partisi atribut A

|S_i| : jumlah kasus pada partisi ke i

|S| : jumlah kasus dalam S

Sementara itu, perhitungan nilai *entropy* dapat dilihat pada persamaam

2.6 .

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad \dots (2.6)$$

Keterangan :

S : himpunan kasus

A : fitur

N : jumlah

P_i : proporsi dari S_i terhadap S

Pseudocode C4.5

Form Tree(T)

- (1) ComputeClassFrequency(T);
 - (2) If OneClass or FewCases
 Return a leaf;
 Create a decision node N;
 - (3) ForEach Attribute A
 ComputeGain(A);
 - (4) N.test = AttributeWithBestGain;
 - (5) If N.test is continuous
 Find Threshold;
 - (6) ForEach T' in the splitting of T
 - (7) If T' is empty
 Child of N is a leaf
 Else
 - (8) Child of N = FormTree(T')
 - (9) Compute Errors of N;
- Return N

I. Perhitungan Akurasi

Perhitungan akurasi dilakukan untuk mengetahui tingkat akurasi dari hasil klasifikasi, dengan cara menghitung jumlah record uji yang kelasnya diprediksi secara tepat. Dapat dilihat pada persamaan 2.7 berikut.

$$\text{Akurasi} = \frac{\text{jumlah prediksi benar}}{\text{jumlah total prediksi}} \times 100\% \quad \dots (2.7)$$

Jumlah prediksi benar adalah jumlah record data uji yang diprediksi kelasnya menggunakan metode klasifikasi dan hasilnya sama dengan kelas sebenarnya. Sedangkan jumlah total prediksi adalah jumlah keseluruhan record yang diprediksi kelasnya (seluruh data uji). Metode klasifikasi berusaha untuk mencari model yang memiliki tingkat akurasi yang tinggi ketika model tersebut diterapkan pada data uji. (sarkar dan Leong, 2000)[9].

J. Penelitian yang Serupa

Ada beberapa penelitian terdahulu yang meneliti tentang klasifikasi rekam medis antara lain :

1. ***Prediksi Status Keaktifan Studi Mahasiswa dengan algortima C5.0 dan K-Nearest Neighbor.*** Oleh Iin Ernawati dari Pascasarjana Institut Pertanian Bogor, Bogor, 2008. Penelitian ini bertujuan untuk membantu menemukan karakteristik mahasiswa aktif maupun tidak

aktif pada sebuah fakultas di sebuah perguruan tinggi swasta di Jakarta Selatan sehingga dapat digunakan sebagai prediksi status studi dimasa datang dengan menggunakan dua metode algoritma K-Nearest Neighbor dan C5.0 kemudian membandingkan kedua metode tersebut.[10]

2. ***Data Mining Kemampuan Siswa berbasis NeuroFuzzy***, Oleh Imron Rosyidi, Mochammad Hariadi, I Ketut Eddy Purnama, Teknik Elektro, FTI,ITS. Kampus ITS Sukolilo, Surabaya, 2011. Penelitian ini dilakukan untuk memprediksi nilai UN setiap siswa sebelum UN berlangsung. Berdasarkan hasil prediksi nilai UN dapat dilakukan langkah-langkah strategis untuk meminimalisir ketidaklulusan siswa dalam UN. Proses Komputasi menggunakan algoritma Neuro Fuzzy. Hasil Prediksi diambil prediksi dengan akurasi paling tinggi dengan waktu komputasi paling pendek. Hasil penelitian ini adalah Neuro Fuzzy dapat digunakan untuk memprediksi UN yang sebenarnya per siswa pada tahun mendatang dengan rata-rata kesalahan 0,568 untuk Bahasa Indonesia, 0,513 untuk matematika, 0,340 untuk Bahasa Inggris. [11]

3. ***Perbandingan metode Nearest Neighbor dan algoritma C45 untuk menganalisis kemungkinan pengunduran diri calon mahasiswa di STIMIK AMIKOM Yogyakarta***, oleh Kusrini, Sri Hartati, Retantyo Wardoyo, Agus Harjoko dari STIMIK AMIKOM Yogyakarta, 2008. Penelitian ini menitikberatkan perbandingan

klasifikasi dan tingkat akurasi dari algoritma k-nearest Neighbor dan algoritma C4.5 untuk pengambilan keputusan dalam proses penjurian calon mahasiswa baru di STMIK AMIKOM Yogyakarta.[12]

4. ***Penerapan Algoritma Modified K-Nearest Neighbor Mknn Untuk Mengklasifikasi Letak Protein pada Bakteri E-Coli, 2010.***

Penelitian ini menekankan pada penerapan Modified K-Nearest Neighbor (MKNN) untuk mengklasifikasikan letak protein pada bakteri E-Coli serta bagaimana pengujian akurasi algoritma Modified K-Nearest Neighbor (MKNN) terhadap banyak tes dan nilai tetangga.[13].

K. Kerangka Pikir

