Analisa Nilai Lamda Model Jarak Minkowsky Untuk Penentuan Jurusan SMA (Studi Kasus di SMA Negeri 2 Tualang)

Khairul Umam Syaliman bin Lukman[#1], Ause Labellapansa[*2]

[#*]*Teknik Informatika, Universitas Islam Riau*
Jl. Kaharudin Nasution no.113 Pekanbaru
[1]khairul.q14@gmail.com
[2]ause.labella@eng.uir.ac.id

Abstract —**SMA Negeri 2 (SMAN 2) is located in Tualang. So far the data report student majors only stored in a database as a final report. Data from the report of the majors could be used as guidelines to determine the students' decision majors for the following year. To take advantage of the data stored in that particular database, we can use data mining disciplines. The method used to make the determination of students majoring done by using Nearest K-Nearest Neighbor (K-NN) algorithm. On the other hand, the method for calculating the distance between the data used models Minkowsky distance with a value of lambda (λ) as a parameter. Lambda values that were analyzed were lambda 1, 2 and 3. Lambda with the value of 1 can generate increasing accuracy in the 11th experiment or with a large amount of data equal to 276 data. Lambda 2 will produce increasing accuracy by the 16th experiment or with the number of training data equal to 356 data while lambda 3 can also produce accuracy continuously increasing by the 11th experiement or with the amount of training data equal to 276 data. The accuracy of the lambda value of 1 is better than lambda 2 and lambda 3. This was proven in 25 experiments at lambda 1 which produces the highest accuracy value for 20 times.**

Keywords — **Classification, Data Mining, K-Nearest Neighbor, Lamda (λ), Minkowsky.**

I. PENDAHULUAN

A. Latar Belakang

Pada saat seorang siswa berada pada kelas XI (Dua SMA), siswa tersebut dihadapi dengan sebuah persoalan baru, yaitu penentuan bidang jurusan. Biasanya penentuan bidang jurusan siswa ini ditentukan dan ditetapkan oleh sekolah yang biasanya dipertimbangkan berdasarkan nilai akademik siswa.

Sekolah Menengah Atas Negeri (SMAN) 2 Tualang merupakan Sekolah Menengah Atas (SMA) yang berada di Tualang. Selama ini di SMAN 2 Tualang data hasil laporan penjurusan siswa yang tersimpan dari tahun ke tahun semakin bertambah, namun data tersebut hanya dijadikan sebagai laporan akhir pada sebuah basis data. Tentu ini menjadi hal yang mubazir mengingat dalam dunia pendidikan proses penentuan bidang jurusan siswa juga menjadi hal yang penting untuk diputuskan. Seharusnya data yang besar tersebut bisa dijadikan pedoman untuk menentukan keputusan jurusan siswa pada tahun berikutnya.

Untuk memanfaatkan data yang tersimpan di basis data tersebut dapat menggunakan disiplin ilmu *data mining*. Munculnya *data mining* didasarkan pada kenyataan bahwa jumlah data yang tersimpan dalam basis data yang semakin besar tanpa ada pemanfaatan lebih lanjut. *Data mining* berusaha menemukan suatu informasi baru yang berguna dengan menggunakan suatu metode.

Dari sekian banyaknya metode yang ada, K-Nearest Neighbor (K-NN) menjadi metode yang memiliki daya tarik tersendiri. K-NN merupakan teknik klasifikasi yang sederhana dengan prinsip memilih tetangga terdekat. Oleh karena itu model pengukuran jarak menjadi penting untuk dipertimbangkan. Ada banyak model pengukuran jarak diantaranya yaitu jarak Euclidean, Manhattan / *city block,* dan Minkowsky.

Penelitian ini menggunakan model pengukuran jarak Minkowsky dengan nilai lamda (λ) 1, 2, dan 3. Minkowsky adalah generalisasi dari model pengukuran jarak Euclidean dan Manhattan. Lamda merupakan parameter penentu dalam model pengukuran jarak Minkowsky ini. Bila lamda bernilai 1maka ruang jarak pada Minkowsky sama dengan Manhattan, dan apabila lamda bernilai 2 maka ruang jaraknya sama dengan Euclidean. Baik Euclidean maupun Manhattan mempunyai kelebihan masing-masing. Euclidean cocok untuk menentukan jarak terdekat (lurus) antara dua data, sedangkan Manhattan sangat teguh untuk mendeteksi *outlier* pada data [4]

Berdasarkan permasalahan di atas, maka penelitian ini membandingkan model pengukuran jarak Minkowsky dengan nilai lamda 1, 2, dan 3 pada kasus penentuan jurusan siswa SMAN 2 Tualang Kabupaten Siak yang bertujuan membandingkan parameter jarak manakah yang paling efisien terhadap kasus penentuan jurusan siswa pada SMAN 2 Tualang.

B. Identifikasi Masalah

Adapun identifikasi masalah dapat diuraikan adalah:

- Kurangnya pemanfaatan data hasil seleksi jurusan siswa yang tersimpan pada basis data di SMAN 2 Tualang.
- Menentukan lamda yang paling efisien pada model jarak Minkowsky dalam kasus penentuan jurusan siswa SMAN 2 Tualang Kabupaten Siak dengan cara melakukan 25 kali percobaan dengan 500 data terhadap lamda 1, 2, dan 3.

C. *Batasan Masalah*

Adapaun batasan permasalahan dalam penelitian ini adalah:

- Penyeleksian Jurusan hanya dilakukan pada siswa SMAN 2 Tualang Kabupaten Siak.
- Untuk membandingkan parameter jarak manakah yang paling efisien maka digunakan model jarak Minkowsky dengan nilai lamda 1, 2, 3 dan tidak mempertimbangkan masalah *outlier* yang mungkin terjadi
- Untuk meneliti hasil jarak Minkowsky, maka dietapkan nilai K yaitu 11 sehingga nilai K tidak berubah-rubah disetiap percobaan lamda 1,2 dan 3
- Hanya menganalisa nilai lamda terhadap tingkat akurasi yang dihasilkan terhadap data latih dan data uji.
- Data yang digunakan dalam penyeleksian jurusan adalah data siswa dari nilai semester 2 kelas X, yaitu nilai IPA (terdiri dari Matematika, Fisika, Kimia, dan Biologi), nilai IPS (terdiri dari Matematika, Ekonomi, Geografi, dan Sosiologi), dan nilai IQ.

D. Tinjauan Pustaka

Obbie [10] melakukan klasifikasi untuk menentukan penjurusan IPA atau IPS siswa SMAN 6 Semarang dengan menggunakan algoritma ID3. Dari hasil penelitian diperoleh peminatan diri sebagai *root* dan dari 371 *dataset* serta 20 data uji diperoleh akurasi sebesar 80%.

Senada dengan Obbie [10], Yeni [11] melakukan penentuan jurusan dengan menggunakan metode KNN dan SMART (*Simple Multi Attribute Rating Technique*). Diperoleh hasil akurasi sebesar 62,5% dengan kriterianya yaitu rata-rata nilai raport semester 1 dan 2 pada 12 mata pelajaran yaitu Matematika, Fisika, Kimia, Biologi, Ekonomi, Geografi, Sejarah, Sosiologi, Bahasa Indonesia, Bahasa Inggris, Seni Budaya dan Bahasa Jepang, hasil tes psikologi, minat siswa, dan saran/ anjuran orang tua.

Rao [8] melakukan pengenalan wajah dengan menggunakan 3 (tiga) fitur lokal yang berbeda yaitu *manhattan distance, weighted angle distance* dan *minkowski distance*. Hasil penelitian menunjukan *minkowski distance* memberikan hasil yang lebih baik.

Heryansyah [1] membuat penelitian yang bertujuan untuk mempermudah penentuan jurusan siswa-siswi SMA di kota Semarang yang bertujuan khususnya bagi guru dan siswa di sekolah tersebut dalam menentukan jurusan yang tepat untuk siswa-siswi SMA sesuai dengan minat serta kemampuan yang dimiliki. Penelitian tersebut menggunakan metode AHP.

Retno [2] membandingkan metode K-NN (model pengukuran jarak Euclidean) dan LDA (*Linear Discriminant Analysis*) dengan variable R-G dan R-G-B dari citra buah belimbing untuk memprediksi tingkat kemanisan buah belimbing. Penelitian tersebut menghasilkan nilai akurasi sebesar 80% dengan metode K-NN nilai variable R-G sedangkan dengan variable R-G-B diperoleh nilai akurasi sebesar 91%. Teknik LDA linear maupun LDA dengan ukuran jarak Mahalanobis menghasilkan akurasi sebesar 91%.

Emerensye [3] telah mengimplementasi algoritma data mining K-Nearest Neighbor (K-NN) dalam pengambilan keputusan pengajuan kredit. Penelitian ini membangun aplikasi untuk mengatisipasi kredit macet karena meningkatnya jumlah pengajuan kredit pada PT. Telkom Kandatel Surabaya Timur dengan menggunakan model pengukuran jarak Euclidean. Pada penelitian ini akan dibuat aplikasi untuk menganalisa model jarak Minkowsky, sehingga akan diperoleh nilai K dan nilai lamda secara otomatis dengan akurasi paling optimal berdasarkan proses training yang telah dilakukan.

II. LANDASAN TEORI

A. Pengertian Data Mining

Munculnya *data mining* didasari pada kenyataan jumlah data yang tersimpan pada basis data yang semakin besar dan hanya dijadikan sebagai laporan akhir tanpa ada pemanfaatan lebih lanjut. Disiplin ilmu data *mining* berusaha memanfaatkan data-data tersebut dengan melakukan suatu proses sehingga menghasilkan suatu informasi, pengetahuan, atau bahkan pola data dari gudang data sehingga dapat dimanfaatkan lebih lanjut.

Menurut Tan [4] defenisi *data mining* adalah sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Menurut David Hand, dkk [5] dari MIT mengartikan bahwa *data mining* adalah analisa terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut. Sedangkan Jiawei Han dan Michelin Kamber [6] menyatakan bahwa *data mining* diungkapkan dengan sederhana mengacu pada ilmu mengenai pemisahan atau "penggalian" pengetahuan dari jumlah data yang besar. Jadi dapat disimpulkan bahwa *data mining* adalah suatu proses untuk menggali pengetahuan, informasi, bahkan pola data yang tersembunyi dari data-data yang tersimpan dalam gudang data.

B. Pekerjaan Dalam Data Mining

Data mining memiliki beragam metode mulai dari klasifikasi, klustering, regresi, seleski variabel dan aturan asosiasi [9]. Klasifikasi sendiri pertama kali diterapkan pada bidang tanaman yang mengklasifikasikan suatu spesies tertentu, seperti yang dilakukan oleh Carolus Linnaeus. Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu berdasarkan karakteristik yang dimiliki [4].

Adapun jenis pekerjaan dalam data *mining* di antaranya adalah:

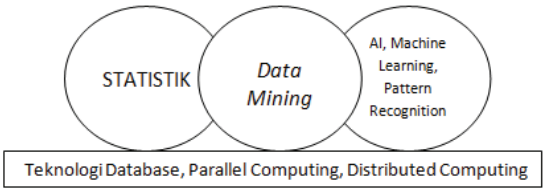
- Deteksi Anomali, berkaitan dengan pengamatan data secara signifikan yang memiliki karakteristik berbeda dari sisa data yang lain. Algoritma deteksi anomali yang baik harus mempunyai laju *error* yang rendah. Deteksi anomali ini dapat diterapkan salah satunya untuk mengetahui pola data yang masuk pada suatu jaringan sehingga pencegahan penyusupan bisa diketahui.
- *Cluster Analysis* atau analisis kelompok melakukan pengelompokan data-data ke dalam sejumlah kelompok (*cluster*) berdasarkan karakteristik masing-masing data. Data-data yang keluar dari batas kesamaan akan terpisah dari kelompoknya, dan data-data yang masuk dalam batas kesamaan suatu kelompok data akan menjadi kelompok data tersebut.
- Analisis Asosiasi, digunakan untuk menemukan pola yang menggambarkan kekuatan fitur dalam data. Penerapan yang paling dekat dalam kehidupan sehari-hari adalah analisis data keranjang belanja. Pembeli adalah ibu rumah tangga, jika ibu tersebut membeli beras, sangat besar kemungkinannya bahwa ibu itu juga akan membeli barang lain, misalnya minyak makan, telur, dan lain sebagainya, sehingga pengecer dapat menentukan barang-barang yang disediakan dalam jumlah yang lebih banyak.

Untuk menggali pengetahuan dari gudang data, setidaknya data akan melalui beberapa tahapan di antaranya adalah [4]:

- *Data Cleaning* atau pembersihan data dari gangguan dan ketidak konsistenan data.
- *Data Integration*, dimana sumber data yang berbeda bisa disatukan.
- *Data Selection*, dimana data yang relevan yang dianalisis dan diterima oleh basis data.
- *Data Transformation*, dimana data diubah atau dikonsolidasikan ke dalam bentuk yang sesuai untuk pertambangan dengan melakukan ringkasan atau agregasi operasi.
- *Data Mining*, proses penting di mana metode cerdas yang diterapkan untuk mengekstrak data.
- *Pattern Evaluation*, untuk mengidentifikasi pola yang benar-benar mewakili pengetahuan didasarkan pada beberapa tingkat kemiripan.
- *Knowledge Presentation*, teknik presentasi atau visualisasi pengetahuan kepada pengguna.

Pada proses satu sampai empat di mana data dipersiapkan untuk digali, pada proses ini *data mining* adalah salah satu langkah dalam seluruh proses penggalian pengetahuan, meskipun data *mining* menemukan pola tersembunyi dari basis data.

Karena keunikan *data mining* inilah para ahli berusaha menentukan posisi bidang *data mining* diantara bidang-bidang ilmu lainnya. Hal ini dikarenakan ada kesamaan antara sebagian bahasan dalam *data mining* dengan bahasan disiplin ilmu lainnya, meskipun tidak seratus persen sama. Salah satunya adalah kesamaan bidang *data mining* dengan bidang statistik yang terletak pada penyampelan, estimasi, dan pengujian hipotesis. Kesamaan dengan kecerdasan buatan (*artificial intelligence*), terletak pada pengenalan pola (*pattern recognition*), dan pembelajaran mesin (*mechine learning*) terletak pada algoritma pencarian, teknik pemodelan, dan teori pembelajaran [4],[9].



Gambar 1 Posisi Data Mining Diantara Beberapa Disiplin Ilmu

Dari Gambar 1 dapat kita lihat bahwa teknologi basis data juga mempengaruhi bidang ilmu *data mining*. *S*edangkan teknik komputasi paralel sering digunakan untuk memberikan kinerja pada ukuran set data yang besar, dan pada komputasi terdistribusi digunakan untuk menyelesaikan masalah data yang tidak dapat disimpan disatu tempat.

Aplikasi yang menggunakan *data mining* bertujuan menyelesaikan permasalahan dengan membangun model berdasarkan data yang sudah digali untuk diterapkan pada data lainnya. Secara umum ada dua jenis aplikasi *data mining* [4]

- Model Deskriptif, adalah suatu model yang bertujuan untuk membantu pengguna agar mudah melihat pola-pola dari data yang ada.
- Model Prediksi, bertujuan untuk dapat melakukan pemetaan dari setiap himpunan variable ke setiap target, yang kemudian menggunakan model tersebut untuk menentukan nilai target pada himpunan baru berdasarkan data-data yang telah ada. Salah satu model prediksi adalah klasifikasi dan regresi. Regresi digunakan untuk variable target kontinu sedangkan

klasifikasi biasanya digunakan untuk variable target diskrit.

C. Klasifikasi

Sebuah sistem yang melakukan proses klasifikasi diharapkan dapat menentukan semua target data input dengan benar, namun tidak dapat dimungkiri bahwa kinerja suatu sistem tidak bisa seratus persen benar, sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja klasifikasi dapat dilakukan dengan menggunakan matriks konfusi (*confusion matrix*).

Dengan mengetahui jumlah data yang diklasifikasi secara benar dan salah, dapat diketahui tingkat akurasi serta laju error dari hasil prediksi. Untuk menghitung akurasi dapat menggunakan rumus di bawah ini:

$$Akurasi = \frac{Jumlah\ data\ yang\ terprediksi\ benar}{Jumlah\ prediksi\ yang\ dilakukan}$$

Sedangkan untuk mengukur laju *error* digunakan formula:

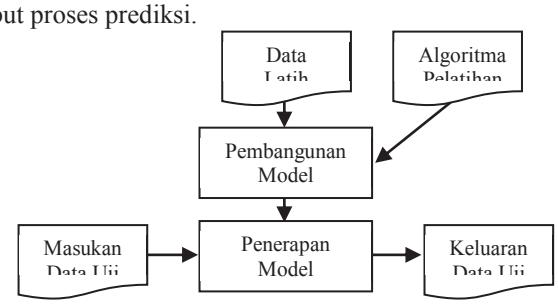
$$Laju\ error = \frac{Jumlah\ data\ yang\ terprediksi\ salah}{Jumlah\ prediksi\ yang\ dilakukan}$$

Dalam [4] klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai prototipe dan penggunaan model tersebut untuk melakukan klasifikasi pada suatu bjek data.

Semua algoritma klasifikasi berusaha membuat model dengan tingkat akurasi tinggi (laju *error* yang rendah). Umumnya, model yang dibangun dapat memprediksi data latih dengan benar, tetapi ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah algoritma klasifikasi ditentukan.

Kerangka kerja klasifikasi meliputi dua langkah proses yaitu induksi yang merupakan langkah untuk membangun model klasifikasi dari data latih yang diberikan dan deduksi merupakan proses untuk menerapkan model tersebut pada data uji sehingga kelas yang sesungguhnya dari data uji dapat diketahui atau biasa disebut proses prediksi.

Gambar 2 merupakan kerangka kerja klasifikasi yang meliputi dua langkah proses, yaitu induksi yang merupakan langkah untuk membangun model klasifikasi dari data latih yang diberikan dan deduksi merupakan proses untuk menerapkan model tersebut pada data uji sehingga kelas yang sesungguhnya dari data uji dapat diketahui atau biasa disebut proses prediksi.



Gambar 2 Proses Kerja Klasifikasi [4]

Ada banyak algoritma pelatihan yang sudah dikembangkan oleh para peneliti, namun berdasarkan cara pelatihannya algoritma ini dapat dibedakan menjadi dua macam, yaitu *eager learner* dan *lazy learner*. *Eager learner* didesain untuk melakukan pembacaan/ pelatihan/ pembelajaran pada data latih agar dapat memetakan dengan benar setiap vektor masukan ke label kelas keluarannya sehingga di akhir proses pelatihannya model sudah dapat memetakan semua vektor data uji ke label kelas dengan benar. Selanjutnya setelah proses pelatihan selesai model disimpan sebagai memori. Proses prediksi dilakukan dengan model yang tersimpan, tidak melibatkan data latih. Cara ini mengakibatkan proses prediksi berjalan dengan cepat, tetapi harus dibayar dengan proses pelatihan yang lama. Algoritma yang masuk dalam kategori ini diantaranya *Artificial Neural Network* (ANN), *Support Vectore Mechine* (SVM), *Decision Tree*, Bayesian, dan lain sebagainya [4]

Sementara Algoritma *lazy learner* adalah algoritma yang masuk dalam kategori sedikit melakukan pelatihan atau sama sekali tidak melakukan pelatihan, algoritma ini hanya menyimpan sebagian atau seluruh data latih yang kemudian menggunakan seluruh atau sebagian dari data latih tersebut untuk proses prediksi. Hal ini mengakibatkan proses prediksi menjadi lama karena model harus membaca kembali data latihnya agar dapat memberikan keluaran label kelas dengan benar pada data uji. Kelebihan algoritma ini adalah proses pelatihan yang berjalan dengan cepat. Algoritma klasifikasi yang termasuk kategori ini di antaranya adalah *K-Nearest Neighboar* (K-NN), *Fuzzy K-Nearest Neighboar* (FK-NN), Regresi Linear, dan sebagainya.

D. K-NN (K-Nearest Neighbor)

Algoritma K-Nearest Neighbor (K-NN) merupakan algoritma yang melakukan proses klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain. [4].

Nilai K pada K-NN merupakan jumlah tetangga terdekat, jika K bernilai 1, maka kelas dari satu data latih yang terdekat akan menjadi kelas bagi data uji yang baru, jika K bernilai 3 akan diambil tiga data latih yang terdekat menjadi kelas untuk data uji yang baru.

Salah satu masalah yang dihadapi K-NN adalah dalam pemilihan nilai K yang tepat. Pemilihan nilai K yang besar dapat mengakibatkan distorsi data yang besar pula. Hal ini dikarenakan setiap tetangga mempunyai bobot yang sama terhadap data uji, sedangkan K yang terlalu kecil bisa menyebabkan algoritma terlalu *sensitive* terhadap *noise*.

K-NN merupakan teknik klasifikasi yang sederhana, tetapi mempunyai hasil kerja yang cukup bagus. Beberapa karakter K-NN adalah sebagai berikut:

- K-NN merupakan algoritma yang menggunakan seluruh atau sebagian data latih untuk melakukan proses klasifikasi. Hal ini mengakibatkan proses prediksi yang sangat lama.

- Algoritma K-NN tidak membedakan setiap fitur (attribut) data denga suatu bobot.
- Hal yang rumit dari K-NN adalah menentukan nilai K yang paling sesuai.
- Karena pada K-NN prinsipnya adalah memilih tetangga terdekat maka model pengukuran jarak juga menjadi hal yang harus diperhatikan.
- K-NN telah memasuki *top ten* algoritma *data mining* [7]

E. Model Pengukuran Jarak

Ada banyak model pengukuran jarak, dan yang paling sering digunakan antara lain model jarak Euclidean, Manhattan, Chebyshev, dan Minkowsky [4].

Pengukuran jarak pada ruang jarak Euclidean menggunakan formula 1:

$$D(x,y) = ||x-y||_2 = \sqrt{(\sum_{j=1}^{N} |x-y|^2} \quad (1)$$

Pengukuran jarak pada ruang jarak Manhattan menggunakan formula 2:

$$D(x,y) = ||x-y||_1 = \sum_{j=1}^{N} |x-y| \quad (2)$$

Pengukuran jarak pada ruang jarak Chebyshev menggunakan formula 3:

$$\mathrm{D}(x,y) = ||x-y||_\lambda = \max\{|x_i - y_i|\} \quad (3)$$

Pengukuran jarak pada ruang jarak Minkowsky menggunakan formula 4:

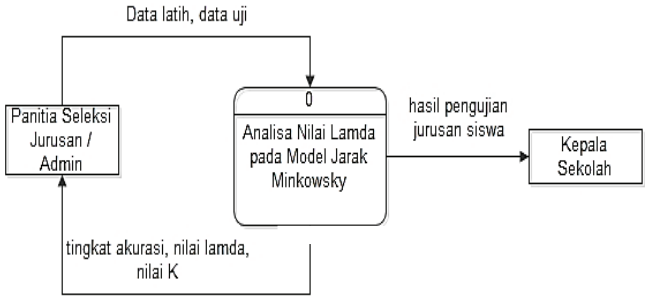
$$D(x,y) = ||x-y||_\lambda = \sqrt[\lambda]{\sum_{j=1}^{N} |x-y|^\lambda} \quad (4)$$

Keterangan:
D = adalah jarak antara data x dan y
N = adalah jumlah fitur (dimensi) data.
Lamda (λ) = adalah parameter jarak minkowksy

Secara umum Minkowsky adalah generalisasi dari Euclidean dan Manhattan. Lamda (λ) merupakan parameter penentu, jika nilai λ = 1 maka ruang jarak Minkowsky sama dengan Manhattan, dan jika λ = 2 ruang jaraknya sama dengan Euclidean dan jika λ= ∞ sama dengan ruang jarak Chebyshev. Setiap model pengukuran jarak mempunyai kelebihan masing-masing, Euclidean cocok untuk menentukan jarak terdekat (lurus) antara dua data, sedangkan Manhattan sangat teguh untuk mendeteksi outlier pada data [4]

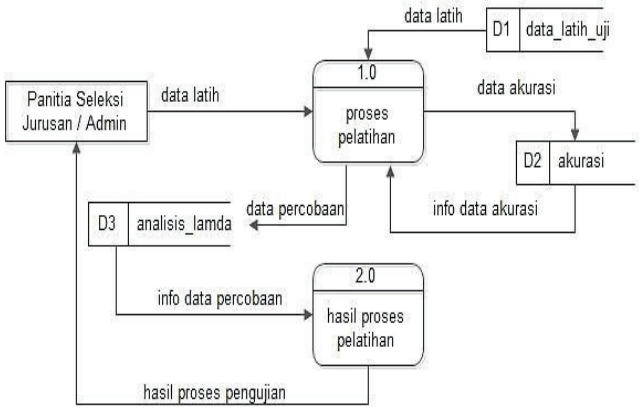
F. Perancangan Sistem

Gambar 3 merupakan *context diagram* dari sistem dimana terdapat dua entitas yaitu staff tata usaha yang bertugas melakukan proses pelatihan dan prediksi jurusan siswa dan kepala sekolah yang nantinya akan menerima hasil pengujian jurusan siswa untuk disetujui.



Gambar 3 *Context Diagram* Sistem

Gambar 4 merupakan DFD level 0 dimana Staff TU akan melakukan proses pelatihan. Proses pelatihan akan menghasilkan nilai akurasi dari setiap subset dan nilai lamda yang paling optimum dalam menghasilkan prediksi. Proses ini menjadi proses yang paling penting karena model yang dihasilkan dari proses pelatihan akan digunakan pada penentuan jurusan. Kemudian Staff TU juga akan melakukan proses pengujian di mana pada proses ini bertujuan untuk melihat sejauh mana sistem dapat memprediksi jurusan siswa di luar data latih yang ada.



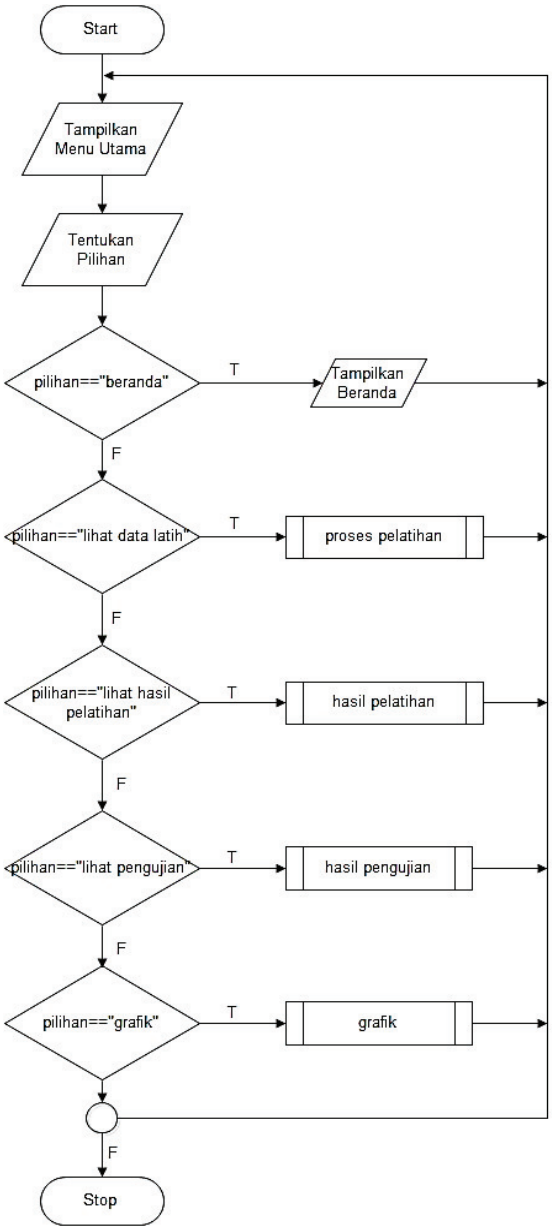
Gambar 4 *Data Flow Diagram* Level *0 Pada Sistem*

III. HASIL DAN PEMBAHASAN

A. Program Flowchart Menu Utama

Pada *flowchart* yang ditampilkan oleh Gambar 5 dapat dilihat bahwa *user* yang menggunakan aplikasi dapat memilih beberapa menu pilihan, diantaranya lihat data latih

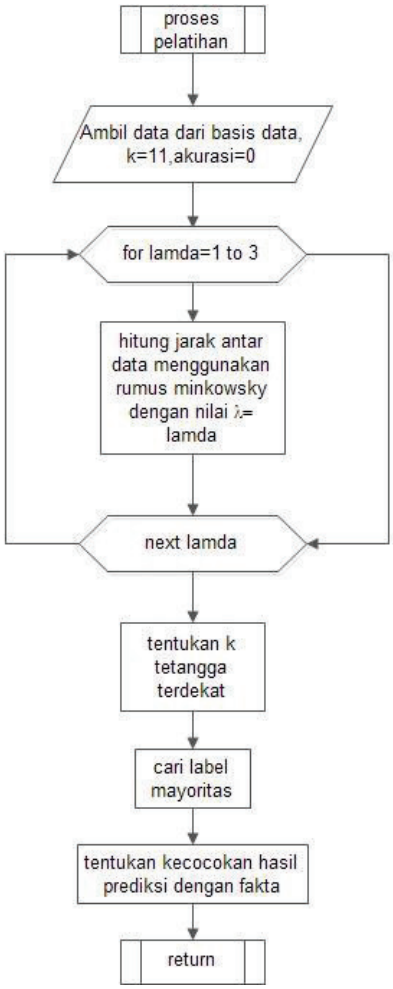
yang akan membawa *user* pada data latih dan dapat melakukan proses pelatihan terhadap data latih yang ada, menu pilihan lihat hasil pelatihan yang akan menampilkan informasi dari pelatihan data sebelumnya, menu pilihan lihat hasil pengujian akan membawa *user* pada hasil pelatihan yang telah dilakukan 25 kali pengujian dengan jumlah data yang berbeda.



Gambar 5 *Program Flowchart* Proses Pelatihan *Data*

B. Proses Pelatihan

Gambar 6 merupakan program flowchart proses pelatihan data dimana hal yang pertama dilakukan pada proses pelatihan data adalah mengambil data dari basis data kemudian melakukan perulangan (looping) untuk nilai lamda = 1 sampai nilai lamda = 3. Dalam proses perulangan ini akan dilakukan pencarian nilai jarak antar data dengan menggunakan rumus Minkowsky dengan nilai λ=lamda. Langkah selanjutnya adalah menetapkan jumlah 11 tetangga terdekat dan mencari kelas mayoritas dari 11 tetangga terdekat tersebut untuk dijadikan kelas dari dari uji dan menentukan kecocokan antara hasil prediksi dengan kenyataan yang dialami pada data tersebut



Gambar 6 *Program Flowchart* Proses Pelatihan *Data*

C. Tahapan Pengujian

Pengujian ini bertujuan untuk menganalisa nilai lamda 1, 2, dan 3 pada metode model jarak minkowsky. Data yang digunakan merupakan data siswa SMA dari nilai semester II kelas X (sepuluh) dengan jumlah data sebanyak 500 data. Adapun kriteria yang dinilai terdiri dari nilai matematika, Fisika, Kimia, Biologi, Ekonomi, Geografi, Sosiologi dan nilai IQ. Sedangkan Class terdiri dari Jurusan IPA atau IPS.

Nilai K yang digunakan adalah K 11. Hal ini dilakukan dengan beberapa alasan yaitu:

- Pemilihan nilai K yang besar dapat mengakibatkan distorsi data yang besar pula sedangkan nilai K yang terlalu kecil dapat menyebabkan algoritma terlalu *sensitive* terhadap *noise*

- Untuk meneliti hasil jarak Minkowsky saat lamda 1,2 dan 3, maka dengan ditetapkan nilai K=11 akan mengakibatkan nilai K tidak berubah-rubah pada setiap percobaan.

Pengujian ini bertujuan untuk menganalisa nilai lamda 1, 2, dan 3 pada metode model jarak Minkowsky. Untuk menganalisa hal tersebut dilakukan percobaan sebanyak 25 kali pelatihan, setiap percobaan dilakukan proses pelatihan terhadap jumlah data yang berbeda, dengan rincian seperti pada Tabel I.

TABEL I
RINCIAN DATA LATIH

Percobaan	Jumlah Data	Kelas			
		IPA	Persentase Jumlah Data	IPS	Persentase Jumlah Data
1	116	61	53%	55	47%
2	132	67	51%	65	49%
3	148	72	49%	76	51%
4	164	75	46%	89	54%
5	180	80	44%	100	56%
6	196	87	44%	109	56%
7	212	97	46%	115	54%
8	228	108	47%	120	53%
9	244	122	50%	122	50%
10	260	132	51%	128	49%
11	276	143	52%	133	48%
12	292	153	52%	139	48%
13	308	159	52%	149	48%
14	324	170	52%	154	48%
15	340	178	52%	162	48%
16	356	188	53%	168	47%
17	372	197	53%	175	47%
18	388	206	53%	182	47%
19	404	216	53%	188	47%
20	420	222	53%	198	47%
21	436	222	51%	214	49%
22	452	222	49%	230	51%
23	468	222	47%	246	53%
24	484	227	47%	257	53%
25	500	230	46%	270	54%

Pengujian pada tahap ini dilakukan dengan melakukan pelatihan pada data latih yang beragam jumlahnya, pelatihan kali ini menjelaskan rincian dari percobaan yang pertama dari 25 kali percobaan (pada Tabel I). Adapun jumlah data yang digunakan berjumlah 116 data, terdiri dari 61 kelompok data IPA dan 55 kelompok data IPS, rincian hasil pelatihan untuk nilai lamda 1, 2, dan 3 terdapat pada Tabel II

TABEL II
RINCIAN HASIL PERCOBAAN PERTAMA

Lamda	Jumlah Data	Prediksi Benar		Prediksi Salah	
		Jumlah	%	Jumlah	%
1	116	100	86.21%	16	13.79%
2	116	100	86.21%	16	13.79%
3	116	99	85.34%	17	14.66%

Pada Tabel II saat lamda bernilai 1, data yang berhasil diprediksi dengan benar berjumlah 100 data dan menghasilkan nilai akurasi 86.21%. Pada lamda 2 nilai akurasi yang dimiliki sama dengan nilai akurasi lamda 1 yaitu 86.21% sedangkan pada lamda 3 nilai akurasi yang dihasilkan hanya sebesar 85.34% saja dengan data yang terprediksi benar sebanyak 99 data.

Selanjutnya dilakukan lagi pelatihan yang menjelaskan proses percobaan kelima yang terdapat pada Tabel I. Data yang digunakan sebanyak 180 data, terdiri dari 80 kelompok data IPA dan 100 kelompok data IPS

TABEL III
RINCIAN HASIL PERCOBAAN KELIMA

Lamda	Jumlah Data	Prediksi Benar		Prediksi Salah	
		Jumlah	%	Jumlah	%
1	180	156	86.67%	24	13.33%
2	180	154	85.56%	26	14.44%
3	180	148	82.22%	32	17.78%

Pada Tabel III, dari 180 data untuk lamda 1 data yang dapat diprediksi dengan benar sebanya 156 data dan mendapatkan nilai akurasi sebesar 86.67%, lamda 2 mendapatkan nilai akurasi sebesar 85.56% dengan data yang dapat diprediksi dengan benar sebanyak 154 data dan lamda 3 dapat memprediksi 148 data dengan benar sehingga hanya menghasilkan akurasi sebesar 82.22%.

Tabel IV merupakan percobaan data yang kesepuluh pada Tabel I. Banyaknya data yang digunakan adalah 260 data, terdiri dari 132 kelompok data IPA dan 128 kelompok data IPS. Dari 260 data yang dilakukan untuk proses pelatihan untuk lamda 1, data yang dapat diprediksi dengan benar sebanyak 230 data dengan nilai akurasi sebesar 88.46%, lamda 2 mendapatkan nilai akurasi sebesar 87.69% dengan data yang dapat diprediksi dengan dengan benar sebanyak 228 data, dan lamda 3 dapat memprediksi 226 data dengan benar dan menghasilkan akurasi sebesar 86.92%.

TABEL IV
RINCIAN HASIL PERCOBAAN KESEPULUH

Lamda	Jumlah Data	Prediksi Benar		Prediksi Salah	
		Jumlah	%	Jumlah	%
1	260	230	88.46%	30	11.54%
2	260	228	87.69%	32	12.31%
3	260	226	86.92%	34	13.08%

Tabel V merupakan percobaan yang kedua puluh dari Tabel I. Pelatihan dilakukan terhadap 420 data, terdiri dari 222 kelompok data IPA dan 198 kelompok data IPS. Dari 420 data yang dilakukan proses pelatihan, saat lamda bernilai 1 nilai akurasi yang didapat sebesar 92.14% dengan jumlah prediksi data yang benar sebanyak 387 data, lamda 2 memiliki nilai akurasi sebesar 91.90% dengan jumlah data yang berhasil diprediksi dengan benar sebanyak 386 data,

lamda 3 berhasil memprediksi data dengan benar sebanyak 385 data dan memberikan nilai akurasi sebesar 91.67%.

TABEL V
RINCIAN HASIL PERCOBAAN KEDUA PULUH

Lamda	Jumlah Data	Prediksi Benar		Prediksi Salah	
		Jumlah	%	Jumlah	%
1	420	387	92.14%	33	7.86%
2	420	386	91.90%	34	8.10%
3	420	385	91.67%	35	8.33%

Percobaan terakhir pada Tabel VI dilakukan terhadap 500 data latih, terdiri dari 230 kelompok data IPA dan 270 kelompok data IPS. Dari percobaan ini diperoleh nilai akurasi yang tertinggi terjadi saat lamda bernilai 1 dan lamda 2, yaitu sebesar 93.20% dengan data yang dapat diprediksi dengan benar sebanyak 466 data, sedangkan lamda 3 mendapatkan tingkat akurasi paling rendah sebesar 92.40% dengan berhasil memprediksi data dengan benar hanya sebanyak 462 data dari 500 data yang digunakan pada proses pelatihan ini.

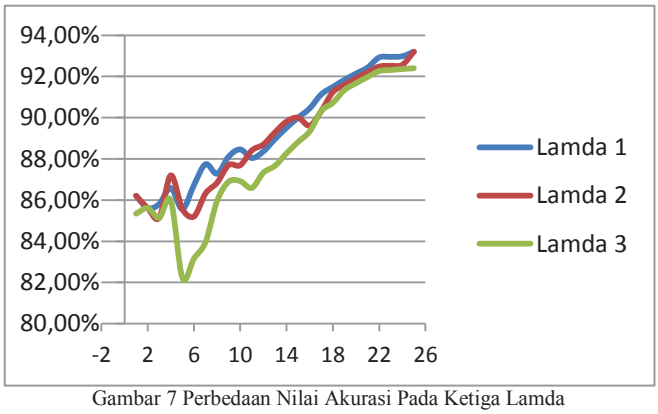
TABEL VI
RINCIAN HASIL PERCOBAAN KEDUA PULUH LIMA

Lamda	Jumlah Data	Prediksi Benar		Prediksi Salah	
		Jumlah	%	Jumlah	%
1	500	466	93.20%	34	6.80%
2	500	466	93.20%	34	6.80%
3	500	462	92.40%	38	7.60%

Tabel VII menjelaskan hasil dari proses pelatihan yang telah dilakukan dari 25 kali percobaan pelatihan dengan menggunakan nilai K=11 untuk setiap nilai lamda dengan Gambar 7 merupakan perbedaan nilai akurasi dari ketiga lamda ini dimana nilai akurasi yang dihasilkan untuk setiap pelatihan dapat berbeda satu dengan yang lainya. Untuk lamda 1 dimulai dari pelatihan ke-11 nilai akurasi yang dihasilkan meningkat untuk setiap pelatihannya sedangkan untuk lamda 2 peningkatan nilai akurasi yang konstan baru terjadi pada percobaan ke-16 dan untuk lamda 3 dimulai dari pelatihan ke-11 nilai akurasi yang dihasilkan juga meningkat, sama halnya dengan lamda 1.

TABEL VII
RINCIAN KESELURUHAN HASIL PERCOBAAN

Percobaan	Jum. Data	Jumlah Prediksi Benar						Keterangan
		λ 1	%	λ 2	%	λ 3	%	
1	116	100	86.21	100	86.21	99	85.34	Lamda 1 & 2
2	132	113	85.61	113	85.61	113	85.61	Lamda 1, 2, 3
3	148	127	85.81	126	85.14	126	85.14	Lamda 1
4	164	142	86.59	143	87.20	141	85.98	Lamda 2
5	180	154	85.56	154	85.56	148	82.22	Lamda 1 & 2
6	196	170	86.73	167	85.20	163	83.16	Lamda 1
7	212	186	87.74	183	86.32	178	83.96	Lamda 1
8	228	199	87.28	198	86.84	196	85.96	Lamda 1
9	244	215	88.11	214	87.70	212	86.89	Lamda 1
10	260	230	88.46	228	87.69	226	86.92	Lamda 1
11	276	243	88.04	244	88.41	239	86.59	Lamda 2
12	292	258	88.36	259	88.70	255	87.33	Lamda 2
13	308	274	88.96	275	89.29	270	87.66	Lamda 2
14	324	290	89.51	291	89.81	286	88.27	Lamda 2
15	340	306	90.00	306	90.00	302	88.82	Lamda 1 & 2
16	356	322	90.45	319	89.61	318	89.33	Lamda 1
17	372	339	91.13	336	90.32	336	90.32	Lamda 1
18	388	355	91.49	354	91.24	352	90.72	Lamda 1
19	404	371	91.83	370	91.58	369	91.34	Lamda 1
20	420	387	92.14	386	91.90	385	91.67	Lamda 1
21	436	403	92.43	402	92.20	401	91.97	Lamda 1
22	452	420	92.92	418	92.48	417	92.26	Lamda 1
23	468	435	92.95	433	92.52	432	92.31	Lamda 1
24	484	450	92.98	448	92.56	447	92.36	Lamda 1
25	500	466	93.20	466	93.20	462	92.40	Lamda 1 & 2



Gambar 7 Perbedaan Nilai Akurasi Pada Ketiga Lamda

Berdasarkan perhitungan, lamda 1, 2, dan 3 bisa menghasilkan nilai akurasi maksimum yang sama. Bila dilihat dari pola grafik yang dihasilkan oleh setiap lamda dengan mengesampingkan tingkat akurasi, maka lamda 1 dan 3 lebih baik dibandingkan dengan lamda 2, karena dari pelatihan ke-11 lamda 1 dan 3 terus mengalami peningkatan nilai akurasi.

Jika yang menjadi prioritas adalah nilai akurasi, maka lamda bernilai 1 yang paling sering menghasilkan akurasi tertinggi. Jadi dapat disimpulkan pada kasus ini, lamda yang bernilai 1 menghasilkan akurasi lebih baik dari lamda bernilai 2 dan lamda yang bernilai 3.

D. Hasil Pengujian

Berdasarkan pengujian yang telah dilakukan, dapat disimpulkan bahwa nilai akurasi pada lamda 1 mulai mengalami peningkatan nilai akurasi pada percobaan ke-11 atau pada jumlah data besar sama dengan 276 data. Lamda 2 baru akan menghasilkan akurasi yang terus meningkat pada percoban ke-16 atau dengan jumlah data besar sama dengan 356 data latih. Terakhir untuk lamda 3, nilai akurasi akan terus meningkat dimulai saat jumlah data besar sama dengan 276 data, atau pada percobaan ke-11.

Dari tingginya tingkat akurasi antar lamda, maka lamda yang bernilai 1 lebih sering menghasilkan akurasi yang tertinggi dibandingkan dengan lamda 2 dan 3, dengan demikian lamda 1 adalah pilihan terbaik untuk menentukan prediksi siswa dalam kasus ini.

Jadi, pada kasus seleksi jurusan siswa di SMAN 2 Tualang Kabupaten Siak dengan sembilan kriteria data dan dua kelas prediksi, akan lebih baik menggunakan nilai lamda 1, selain lebih cepat menghasilkan peningkatan akurasi (yaitu pada banyaknya data besar sama dengan 276 data), lamda 1 juga lebih sering menghasilkan tingkat akurasi yang tertinggi dari pada lamda yang bernilai 2 dan 3.

IV. Simpulan dan Saran Penelitian

A. Simpulan

Berdasarkan hasil analisa baik pada sistem maupun perhitungan secara manual yang dilakukan maka dapat disimpulkan beberapa point berikut ini:

- Telah berhasil membuat aplikasi Analisa Nilai Lamda Pada Model Jarak Minkowsky, dengan nilai lamda yang memiliki tingkat akurasi tertinggi dari 25 kali pengujian yang dilakukan terhadap jumlah data latih yang berbeda adalah nilai lamda 1
- Pada kasus ini, lamda bernilai 1 sudah dapat menghasilkan akurasi yang terus meningkat pada percobaan ke-11 atau dengan jumlah data besar sama dengan 276 data. Lamda 2 baru akan menghasilkan akurasi yang terus meningkat pada percoban ke-16 atau dengan jumlah data latih besar sama dengan 356 data. Sedangkan lamda 3 juga dapat menghasilkan akurasi yang terus meningkat pada percobaan ke-11 atau pada kondisi jumlah data latih besar sama dengan 276 data.
- Nilai akurasi pada lamda 1 lebih baik dari lamda 2 dan lamda 3, terbukti dengan 25 kali percobaan lamda 1 menghasilkan nilai akurasi tertinggi sebanyak 20 kali.
- Dari beberapa simpulan di atas pada kasus penyeleksian jurusan siswa di SMAN 2 Tualang Kabupaten Siak, akan lebih baik menggunakan nilai lamda 1 pada metode model jarak Minkowsky dengan sembilan kriteria data dan dua kelas data.

B. Saran

Adapun saran yang yang dapat diberikan untuk penelitian selanjutnya dalam konteks pengembangan ilmu pengetahuan mengenai *data mining* ini adalah:

- Penelitian berikutnya dapat mengembangkan aplikasi untuk menganalisa nilai lamda 4, 5, dan 6.
- Penelitian berikutnya dapat meneliti pengaruh nilai lamda terhadap jumlah attribut yang melekat pada data.

Daftar Pustaka

[1] Heryansyah Fanoza, Hidayat, Wahyu, Darmawan, Riza Budi. 2013. Aplikasi Sistem Pendukung Keputusan Pemilihan Jurusan siswa-siswi SMA (IPA/IPS/Bahasa) Menggunakan Metode AHP (Studi Kasus SMA di Kota Semarang). http://arumtyas05.com/2013/01/sistem-pendukung-keputusan-penentuan.html. 12 Januari 2014.

[2] Retno Nugroho Whidhiasih, Wahanani, Nursinta Adi., Supriyanto. 2013, *Klasifikasi Buah Belimbing Berdasarkan Citra Red-Green-Blue Menggunakan KNN Dan LDA*, System Embedded & Logic. Vol 1: 29-35.

[3] Emerensye S. Y P,. 2012. Implementasi Algoritma Data Mining K-Nearest Neighbour(K-NN) Dalam Pengambilan Keputusan Pengajuan Kredit. http://eprints.undip.ac.id/39222/1/Emerensye_Sofia.pdf. 20 Oktober 2014.

[4] Eko Prasetyo. 2012. Data Mining-Konsep dan Apliksai Menggunakan MATLAB. Andi Offset. Yogyakarta.

[5] Prabowo Pudjo Widodo, dkk. *Penerapan Data Mining Dengan MATLAB*. Rekayasa Sains. Bandung.

[6] Han Jiawei. 2006. *Data Mining: Concepts and Techniques* Second Edition. Elsevier. San Francisco.

[7] Wu Xindong & Vipin Kumar. 2009. *The Top Ten Algorithms in Data Mining*. Taylor & Francis Group. United States of America.

[8] Rao, M.K., Swamy, K.V., seetha, K.A., dan Mohan, B.C., 2012. *Face Recognition Using Different Local Feature with Different Distance Techniques*, International Journal of Computer Science. Engineering and Information Technology (IJCSEIT), Vol.2, No.1, pp 67-74, DOI: 10.5121/ijcseit.2012.2107.

[9] Budi Santosa. 2007. Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Graha Ilmu. Yogyakarta

[10] Obbie Kristanto. 2014. Penerapan Algoritma Klasifikasi Data Mining ID3 Untuk Menentukan Penjurusan Siswa SMAN 6 Semarang. http://eprints.dinus.ac.id/13334/1/jurnal_14005.pdf

[11] Yeni Kustiyahningsih dan Nikmatus Syafa'ah. 2015. *Sistem Pendukung Keputusan Untuk Menentukan Jurusan Pada Siswa SMA Menggunakan Metode KNN dan SMART*. Jurnal Sistem Informasi Indonesia. Volume 1 Nomor 1