BAB II
KERANGKA TEORITIS

A. Landasan Teori

1. Data Mining

Menurut (Turban, 2001) Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstrasi pengetahuan (*knowledge*) secara otomatis. Definisi lain diantaranya adalah pembelajaran berbasis induksi (*induction – based learning*) adalah proses pembentukan definisi-definisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik dari konsepkonsep yang akan dipelajari. Knowledge Discovery in Databases (KDD) adalah penerapan metode saintifik pada data mining. Dalam konteks ini data mining merupakan satu langkah dari proses KDD.

Menurut (Tan, 2006) KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interpretasi dan visualisasi dari pola-pola sejumlah data. Serangkaian proses tersebut memiliki tahap sebagai berikut :

a. Pembersihan data yaitu untuk membuang data yang tidak konsisten dan noise.
b. Integrasi data yaitu penggabungan data dari beberapa sumber
c. Transformasi data yaitu data diubah menjadi bentuk yang sesuai untuk di mining.
d. Aplikasi teknik data mining yaitu proses ekstraksi pola dari data yang ada.
e. Evaluasi pola yang ditemukan yaitu proses interpretasi pola menjadi pengetahuan yang dapat digunakan untuk mendukung pengambilan keputusan.
f. Presentasi pengetahuan yaitu dengan teknik visualisasi.

Tahap ini merupakan bagian dari proses pencarian pengetahuan yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Langkah terakhir KDD adalah mempresentasikan pengetahuan dalam bentuk yang mudah dipahami pengguna.

2. Clustering

Menurut (Tan, 2006) *Clustering* adalah sebuah proses untuk mengelompokan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum.

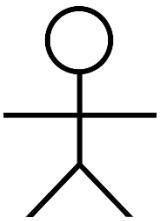
Dengan menggunakan *clustering* ini, kita dapat mengkalsifikasikan daerah yang padat, menemukan pola-pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut data. Dalam data mining, usaha difokuskan pada metode-metode penemuan untuk cluster pada basis data berukuran besar secara efektif dan efisien. Beberapa kebutuhan clustering dalam data mining meliputi skalabilitas, kemampuan untuk menangani tipe atribut yang berbeda mampu menangani dimensionalitas yang tinggi, menangani data yang mempunyai noise, dan dapat diterjemakan dengan mudah.

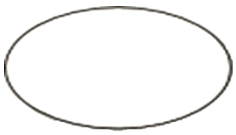
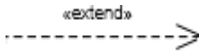
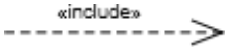
Adapun tujuan dari data clustering ini adalah untuk meminimalisasikan objektif function yang di-set dalam proses clustering, yang pada umumnya berusaha meminimalisasikan variasi dalam suatu cluster. Dan meminimalisasikan variasi antar cluster. Secara garis besar, terdapat beberapa metode klasifikasi data. Pemilihan metode clustering tergantung pada tipe data dan tujuan *clustering* itu sendiri.

3. Unified Modeling Language (UML)

Menurut (Sri Dharwiyanti, 2003) *Unified Modeling Language* (UML) adalah sebuah bahasa yang berdasarkan grafik/gambar untuk memvisualisasi, menspesifikasikan, membangun, dan pendokumentasian dari sebuah sistem pengembangan *software* berbasis OO (*Object Oriented*).

Tabel 2. 1 Simbol Unified Modeling Language (UML)

Simbol	Keterangan
	*Actor* Orang proses, atau sistem lain yang berinteraksi dengan sistem informasi yang akan dibuat di luar sistem informasi yang akan dibuat itu sendiri, jadi walaupun simbol dari actor adalah gambar orang, biasanya

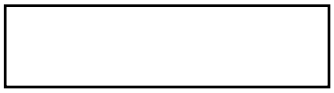
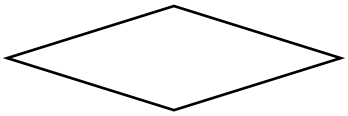
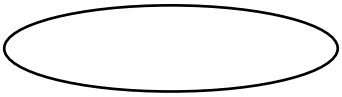
Simbol	Keterangan
	dinyatakan menggunakan kata benda di awal frase nama actor.
	*Use Case* Fungsionalitas yang disediakan sistem sebagai unit-unit yang saling bertukar pesar antar unit atau actor biasanya dinyatakan dengan menggunakan kata kerja di awal frase nama use case.
	Association Komunikasi antara actor dan use case yang berpartisipasi pada use case atau use case memiliki interraksi dengan actor.
	*Extend* Relasi use case tambahan ke sebuah use case dimana use case yang ditambahkan dapat berdiri sendiri walau tanpa use case tambahan memiliki nama depan yang sama dengan use case yang di tambahkan.
	*Generalization* Hubungan generalisasi dan spesialisasi (umum-khusus) antara dua buah use case dimana fungsi yang satu adalah fungsi yang lebih umum dari lainnya.
	*Include* Relasi use case tambahan ke sebuah use case dimana use case yang ditambahkan memerlukan use case ini untuk menjalankan fungsional atau sebagai syarat dijalankan use case ini.

4. Entity Realtionship Diagram (ERD)

Menurut (Sutanta, 2009) dari buku "Basis Data Dalam Tinjauan Konseptual", *Entity Relationship Diagram* (ERD) merupakan suatu model data yang dikembangkan berdasarkan objek. *Entity Relationship Diagram* (ERD) digunakan untuk menjelaskan hubungan antar data dalam basis data kepada pengguna secara logis.

Entity Relationship Diagram (ERD) didasarkan pada suatu persepsi bahwa real world terdiri atas obyek-obyek dasar tersebut. Penggunaan *Entity Relationship Diagram* (ERD) relatif mudah dipahami, bahkan oleh para pengguna yang awam. Bagi perancang atau analis sistem, *Entity Relationship Diagram* (ERD) berguna untuk memodelkan sistem yang nantinya, basis data akan di kembangkan. Model ini juga membantu perancang atau analis sistem pada saat melakukan analis dan perancangan basis data karena model ini dapat menunjukkan macam data yang dibutuhkan dan kerelasian antar data di dalamnya.

Tabel 2. 2 Simbol ERD (Entity Relationship Diagram)

Simbol	Keterangan
	Entitas, yaitu kumpulan dari objek yang dapat diidentifikasikan secara unik
	Relasi, yaitu hubungan yang terjadi antara salah satu lebih entitas. Jenis hubungan antara lain. *One to one*, *one to many*, dan *many to many*
	Atribut, yaitu karakteristik dari entitas atau relasi yang merupakan penjelasan detail tentang entitas
	Hubungan antara entitas dengan atributnya dan dan himpunan entitas dengan himpunan relasinya.

5. Sekolah

Menurut (Atmodiwiro, 2000) Sekolah adalah sistem interaksi sosial suatu organisasi keseluruhan terdiri atas interaksi pribadi terkait bersama dalam suatu hubungan organic Sedangkan berdasarkan undang-undang no 2 tahun 1989 sekolah adalah satuan pendidikan yang berjenjang dan berkesinambungan untuk menyelenggarakan kegiatan belajar mengajar.

Menurut (Daryanto, 1997) sekolah adalah bangunan atau lembaga untuk belajar serta tempat menerima dan memberi pelajaran. Jadi, sekolah sebagai suatu sistem sosial dibatasi oleh sekumpulan elemen kegiatan yang berinteraksi dan membentuk suatu kesatuan sosial sekolah yang demikian bersifat aktif kreatif artinya sekolah dapat menghasilkan sesuatu yang bermanfaat bagi masyarakat dalam hal ini adalah orang-orang yang terdidik.

Dari definisi tersebut bahwa sekolah adalah suatu lembaga atau organisasi yang diberi wewenang untuk menyelenggarakan kegiatan pembelajaran. Sebagai suatu organisasi sekolah memiliki persyaratan tertentu.

Menurut (Arbi, 1990) dikutip dari buku yang ditulis oleh (Pidarta, 1997) sekolah adalah suatu lembaga atau tempat untuk belajar seperti membaca, menulis dan belajar untuk berperilaku yang baik. Sekolah juga merupakan bagian integral dari suatu masyarakat yang berhadapan dengan kondisi nyata yang terdapat dalam masyarakat pada masa sekarang. Sekolah juga merupakan lingkungan kedua tempat anak-anak berlatih dan menumbuhkan kepribadiannya.

6. Pengembangan *System Development Life Cycle* (SDLC)

Menurut (Raymond McLeod Jr, 2007) Pendekatan sistem merupakan sebuah metodologi. Metodologi adalah satu cara yang direkomendasikan dalam melakukan sesuatu. Pendekatan sistem adalah metodologi dasar dalam memecahkan segala jenis masalah. Siklus hidup pengembangan sistem (*System Development Life Cycle* – SDLC) adalah aplikasi dari pendekatan sistem bagi pengembangan suatu sistem informasi.

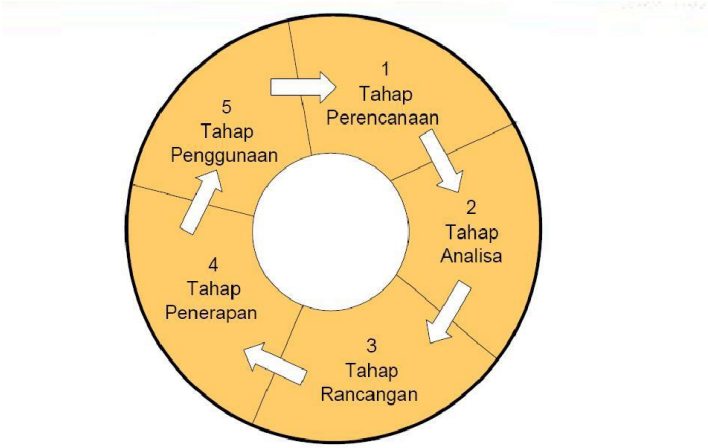
Terdapat beberapa tahapan pekerjaan pengembangan yang perlu dilakukan jika suatu proyek ingin memiliki kemungkinan berhasil yang besar. Tahapan-tahapan tersebut adalah:

a. Perencanaan
b. Analisis
c. Desain
d. Implementasi
e. Penggunaan

Proyek dan sumber daya yang dibutuhkan untuk melakukan pekerjaan direncanakan kemudian disatukan. Sistem yang ada juga dianalisis untuk memahami masalah dan menentukan persyaratan fungsional dari sistem

yang baru. Sistem baru ini kemudian dirancang dan diimplementasikan. Setelah implementasi, sistem kemudian digunakan, idealnya untuk jangka waktu yang lama.

Karena pekerjaan-pekerjaan di atas mengikuti satu pola yang teratur dan dilaksanakan dengan cara dari atas ke bawah, SDLC tradisional sering kali disebut sebagai pendekatan air terjun (waterfall approach). Aktivitas ini memiliki aliran satu arah menuju ke penyelesaian proyek.



Gambar 2. 1 Pola Melingkar dari Siklus Hidup Sistem

(Raymond McLeod Jr, 2007)

Ketika sebuah sistem telah melampaui masa manfaatnya dan harus diganti, satu siklus hidup baru akan dimulai dengan diawali oleh tahap perencanaan. Mudah bagi kita untuk melihat bagaimana SDLC tradisional dapat dikatakan sebagai suatu aplikasi dari pendekatan sistem. Masalah akan didefinisikan dalam tahap-tahapan perencanaan dan analisis. Solusi-solusi alternatif diidentifikasi dan dievaluasi dalam tahap desain. Lalu, solusi yang terbaik diimplementasikan dan digunakan. Selama tahap penggunaan, umpan balik dikumpulkan untuk melihat seberapa baik sistem mampu memecahkan masalah yang telah ditentukan.

7. Bahas Pemrograman

a. Hypertext Preprocessor (PHP)

Menurut (Anhar, 2010) “PHP singkatan dari Hypertext Prepocessor yaitu bahasa pemrograman web server-side yang bersifat open source. PHP (Hypertext Preprocessor) adalah salah satu bahasa

pemrograman open source yang sangat cocok atau dikhususkan untuk pengembangan web dan dapat ditanamkan pada sebuah skripsi HTML. Bahasa PHP dapat dikatakan menggambarkan beberapa bahasa pemrograman seperti C, Java, dan Perl serta mudah untuk dipelajari. Adapun pengertian lain PHP adalah akronim dari Hypertext Preprocessor, yaitu suatu bahasa pemrograman berbasiskan kode-kode (script) yang digunakan untuk mengolah suatu data dan mengirimkannya kembali ke web browser menjadi kode HTML.

b. Hypertext Markup Language (HTML)

Menurut (Janner, 2010), HTML adalah bahasa markup untuk menyebarkan informasi pada web. Ketika merancang HTML, ide ini diambil dari Standart Generalized Markup Language (SGML). SGML adalah cara yang terstandarisasi dari pengorganisasian dan informasi yang terstruktur di dalam dokumen atau sekumpulan dokumen. Walaupun HTML tidak dengan mudah dapat dipahami kebanyakan orang, ketika diterbitkan penggunaannya menjadi jelas. HTML adalah singkatan dari Hypertext Markup Language yaitu bahasa pemrograman standar yang digunakan untuk membuat sebuah halaman web, yang kemudian dapat diakses untuk menampilkan berbagai informasi di dalam sebuah penjelajah web internet (browser).

8. Database Yang Digunakan

a. Database

Menurut (Mustakini, 2009) database adalah kumpulan dari data yang saling berhubungan satu dengan yan lainnya, tersimpan di perangkat keras komputer dan digunakan perangkat lunak untuk memanipulasi.

b. MySQL

Menurut (Mustakini, 2009) “My SQL adalah sebuah basis data yang mengandung satu atau jumlah tabel. Tabel terdiri atas sejumlah baris dan setiap baris mengandung satu atau sejumlah tabel. Tabel terdiri atas sejumlah baris dan setiap baris mengandung satu atau sejumlah tabel”.

Tipe data My SQL (Kustiyahningsih, 2011) adalah data yang terdapat dalam sebuah tabel berupa field-field yang berisi nilai dari data tersebut dan nilai data dalam field memiliki tipe sendiri-sendiri.

9. Web Server

Menurut (MCool, 2018) Apache adalah sebuah nama web server yang bertanggung jawab pada request-response HTTP dan logging informasi secara detail (kegunaan basicnya). Selain itu, Apache juga diartikan sebagai suatu web server yang kompak, modular, mengikuti standar protokol HTTP, dan tentu saja sangat digemari.

10.Intranet

Menurut (Prakoso, 2007) Intranet adalah sebuah kumpulan jaringan komputer lokal yang menggunakan perangkat lunak internet dan protokol TCP/IP atau HTTP. Oleh karena itu, sebuah jaringan intranet memiliki semua fasilitas yang dimiliki oleh internet seperti e-mail, File Transfer Protocol (FTP), dan lain sebagainya. Jaringan intranet merupakan jaringan internet yang hanya dimiliki oleh perusahaan dan tidak dapat diakses dari luar.

B. Algoritma K-Means

Menurut (Xindong Wu, 2009) Algoritma K-Means merupakan algoritma pengelompokkan iteratif yang melakukan partisi set data ke dalam sejumlah K cluster yang sudah ditetapkan di awal. Algoritma K-Means sederhana untuk diimplementasikan dan dijalankan, relatif cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, K-Means menjadi salah satu algoritma yang paling penting dalam bidang data mining.

Secara historis, bentuk esensial K-Means ditemukan oleh sejumlah peneliti dari lintas disiplin ilmu. yang paling berpengaruh adalah Lloyd (1982), Forgey (1967), Friedman dan Rubin (1967), McQueen (1967). Algoritma K-Means berkembang hingga menjadi konteks yang lebih besar sebagai algoritma hill-climbing, seperti yang disampaikan oleh Gray dan Nuhoff (1998).

K-Means dapat diterapkan pada data yang direpresentasikan dalam r-dimensi ruang tempat. K-Means mengrlompokkan set data r-dimensi, $X = \{x_i | i = 1, \dots, N\}$ dimana $x_i \in \Re^d$ yang menyatakan data ke-i sebagai "titik data". Seperti yang dijelaskan sebelumnya bahwa K-Means mempartisi X ke dalam K cluster, Algoritma K-Means mengelomPokkan semua titik data dalam X sehingga setiap titik x_i hanya jatuh dalam satu dari K partisi. Yang perlu diperhatikan adalah titik berada dalam cluster yang mana dilakukan dengan cam dengan memberikan setiap titik sebuah ID cluster. Titik dengan ID cluster yang sama berarti berada dalam satu cluster yang sama, sedangkan titik dengan ID cluster yang berbeda berada dalam cluster yang berbeda. Untuk

menyatakan hal ini, biasanya dilakukan dengan vektor keanggotaan cluster m dengan panjang N, di mana m, bernilai ID cluster titik x_i.

Parameter yang harus dimasukkan ketika menggunakan algoritma K-Means adalah nilai K. Nilai K yang digunakan biasanya didasarkan pada informasi yang diketahui sebelumnya tentang sebenarnya berapa banyak cluster data yang muncul dalam X, berapa banyak cluster yang dibutuhkan untuk penerapannya, atau jenis cluster dicari dengan mengeksplorasi/melakukan percobaan dengan beberapa nilai K. Berapa nilai K yang dipilih tidak perlu memahami bagaimana K-Means mempartisi set data X.

Dalam K-Means, setiap cluster dari K cluster diwakili oleh titik tunggal dalam $\Re^d$. Set representatif cluster dinyatakan $C = \{c_j | j = 1, \dots, K\}$. Sejumlah K representatif cluster tersebut disebut juga sebagai cluster means atau cluster centroid (atau centroid saja). Untuk set data dalam X dikelompokkan berdasarkan konsep kedekatan atau kemiripan. Meskipun konsep yang dimaksud untuk data – data yang berkumpul dalam satu cluster adalah data – data yang mirip, tetapi kuantitas yang digunakan untuk mengukurnya adalah ketidakmiripan (dissimilarity). Artinya, data-data dengan ketidakmiripan (jarak) yang kecil/dekat maka lebih besar kemungkinannya untuk bergabung dalam satu cluster.

Pada saat data sudah dihitung ketidakmiripan terhadap setiap centroid, maka selanjutnya dipilih ketidakmiripan yang paling kecil sebagai cluster yang akan diikuti sebagai relokasi data pada cluster di sebuah iterasi. Relokasi sebuah data dalam cluster yang diikuti dapat dinyatakan dengan nilai keanggotaan a yang bemilai 0 atau 1. Nilai 0 jika tidak menjadi anggota sebuah cluster dan 1 jika menjadi anggota sebuah cluster. Karena K-Means mengelompokkan secara tegas data hanya pada satu cluster, maka dari nilai a sebuah data pada semua cluster, hanya satu yang bemilai 1, sedangkan lainnya 0 seperti dinyatakan oleh persamaan berikut :

$$a_{ij} = \begin{cases} 1 & arg\min_{j} \{(X_i, C_j)\} \\ 0 & L = lainnya \end{cases}$$

$d(X_i, C_j)$ menyatakan ketidakmiripan (jarak) dari data ke-*i* ke cluster C_j

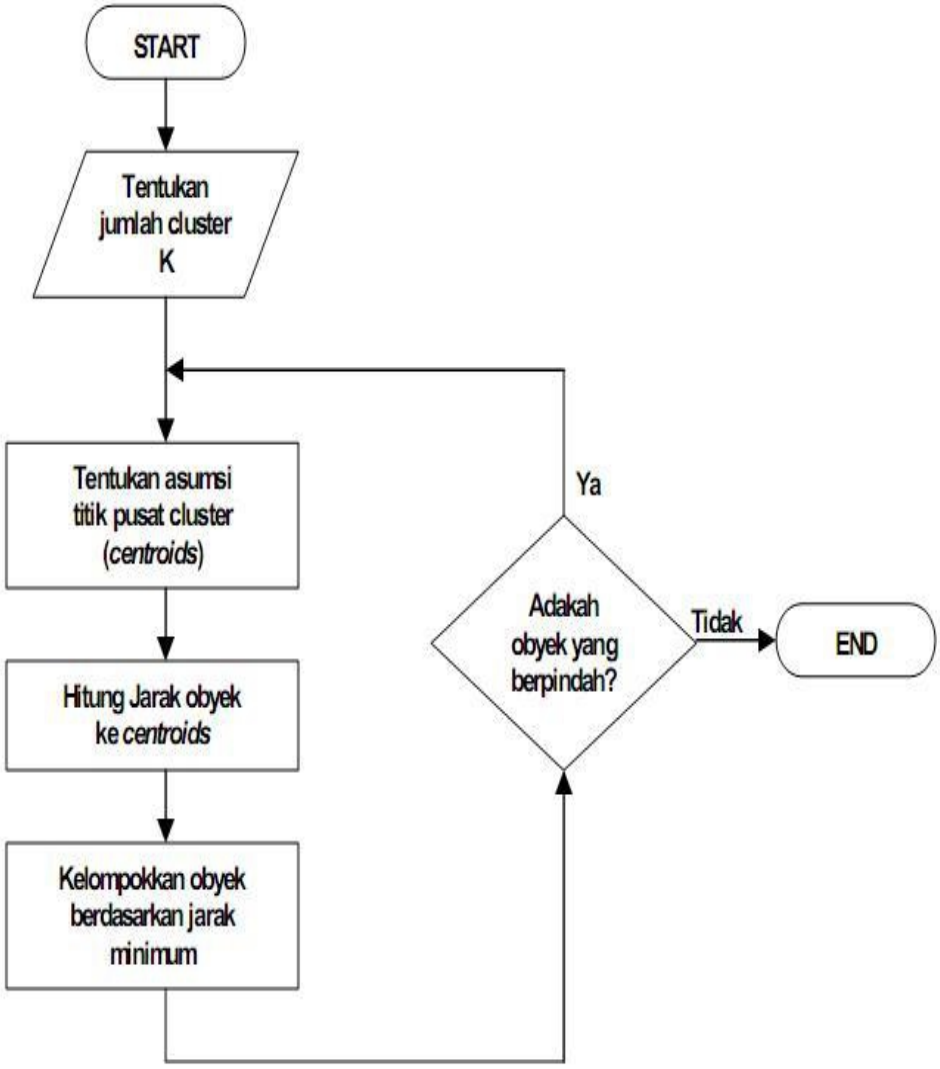
Sementara relokasi centroid untuk mendapatkan titik centroid C didapatkan dengan menghitung rata – rata setiap fitur dari semua data yang tergabung dalam setiap cluster. Rata-rata sebuah fitur dari semua data dalam sebuah cluster dinyatakan oleh persamaan berikut :

$$C_j = \frac{1}{Nk} \sum_{l=1}^{Nk} x_{jl}$$

Nk adalah jumlah data yang tergabung dalam sebuah cluster. Jika diperhatikan dari langkahnya yang selalu memilih cluster terdekat, maka sebenarnya K-Means berusaha untuk meminimalkan fungsi objektif/fungsi biaya non-negatif, seperti dinyatakan oleh persamaan berikut :

$$J = \sum_{i=1}^{N} \sum_{l=1}^{K} a_{ic} d(x_i, C_l)^2$$

Dengan kata lain, K-Means berusaha untuk meminimalkan total jarak kuadrat (*squared distance*) di antara setiap titik x_i dan representasi cluster c_j terdekat. Berikut adalah tahapan dalam Algoritma K-Means :



Gambar 2. 2 Tahapan Algoritma K-Means

Keterangan :

1. Inisiaiisasi adalah tentukan nilai K sebagai jumlah cluster yang diinginkan dan metrik ketidakmiripan (jarak) yang diinginkan. Jika perlu, tetapkan ambang batas perubahan fungsi objektif dan ambang batas perubahan fungsi objektif dan ambang batas perubahan posisi centroid.
2. Pilih K data dari set data X sebagai centroid.
3. Alokasikan semua data ke centroid terdekat dengan metrik jarak yang sudah ditetapkan (memperbarui cluster ID setiap data).
4. Hitung kembali centroid C berdasarkan data yang mengikuti cluster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu (a) perubahan fungsi objektif sudah di bawah ambang batas yang diinginkan; atau (b) tidak ada data yang berpindah cluster; atau (c) perubahan posisi centroid sudah di bawah ambang betas yang ditetapkan.

Algoritma K-Means mengelompokkan set data X dalam langkah iteratif. Berikut dua langkah utamanya, yaitu (1) penentuan kembali ID cluster dan semua titik data dalam X, dan (2) memperbarui representasi cluster (centroid) berdasarkan titik data dalam setiap cluster.

Algoritma bekerja sebagai berikut yaitu pertama, representasi cluster diinisialisasi dengan memilih K data dalam $\Re^d$ secara acak. Selanjutnya, secara iteratif melakukan dua langkah berikut sampai tercapai kondisi konvergen.

Langkah 1 Data assignment. Setiap data ditetapkan ke centroid terdekat dengan pemecahan hubungan apa adanya. Hasilnya berupa data yang terpartisi.

Langkah 2 Relocation of "means". Setiap representasi cluster direlokasi ke pusat (center) dengan rata – rata aritmetika dari semua data yang ditetapkan masuk ke dalamnya. Rasionalnya langkah ini didasarkan pada observasi bahwa dalam memberikan set titik, representasi tunggal yang terbaik untuk set tersebut (dalam hal meminimalkan jumlah lcuadrat jarak Euclidean di antara setiap titik data. Hal ini jugalah yang menyebabkan metode ini sering disebut dengan cluster mean atau cluster centroid, seperti nama yang dimiliki.

Algotitma K-Means tnencapai kondisi konvergen ketika pengalokasian kembali titik data (dan juga lokasi centroid c_j) tidak lagi berubah. Proses dari iterasi ke iterasi hingga dicapai kondisi konvergen juga dapat yang didapatkan. Pada kondisi yang semakin konvergen dapat diamati bahwa dari nilai fungsi objekfif akan semakin menurun. Penyelesaian masalah *local optima* dapat diselesaikan dengan menjalankan

algoritma beberapa kali dengan inisial centroid yang berbeda kemudian memilih hasil yang terbaik.

Pemilihan nilai K yang optimal juga menjadi hal sulit dilakukan. Jika ada informasi mengenai set data, seperti jumlah partisi yang secara alami menggambarkan set data, maka informasi tersebut dapat digunakan untuk memilih nilai K yang optimal. Jika tidak, maka harus menggunakan beberapa kriteria lain untuk memilih K, kemudian menyelesaikan masalah pemilihan model tersebut.

Solusi yang naif adalah dengan mencoba beberapa nilai K berbeda dan memilih clustering yang nilai fungsi objektifinya minimal. Sayangnya nilai yang diberikan oleh fungsi objektif tidak cukup informatif untuk digunakan sebagai harapan penyelesaian masalah ini. Misalnya biaya solusi optimal terhadap peningkatan K menurun sampai menjadi nol ketika jumlah cluster sama dengan jumlah titik data berbeda. Hal tersebut menjadi lebih sulit jika menggunakan fungsi objektif untuk (a) secara langsung membandingkan solusi dengan jumlah cluster berbeda, dan (b) mencari nilai K yang optimal.

Oleh karena itu jika K yang dibutuhkan tidak diketahui, salah satu cara yang biasa digunakan adalah menjalankan K-Means dengan nilai K berbeda dan kemudian menggunakan beberapa kriteria yang lain yang cocok untuk memilih satu hasil yang terbaik. Misalnya X-Means menambahkan istilah kompleksitas (yang meningkat terhadap K) pada fungsi objektif yang asli dan kemudian mengidentifikasi K yang meminimalkan biaya penyesuaian (Pelleg dan Moore, 2000). Alternatif yang lain adalah secara progresif meningkatkan jumlah cluster dengan gabungan kriteria pemberhentian yang cocok. Menurut (Steinbach, 2000) Bisecting K-Means melakukan hal tersebut dengan meletakkan semua data dalam cluster tunggal dan kemudian secara rekursif memecah cluster paling padat menjadi dua cluster menggunakan 2-means.

Pemiiihan K titik data sebagai centroid awal juga memengaruhi basil clustering. Sifat ini menjadi karakteristik alarni K-Means yang dapat mengakibatkan hasil cluster yang didapat pada percobaan berbeda mendapatkan hasil berbeda. Kondisi seperti ini dikenal dengan solusi yang *local optima*, yang artinya algoritma K-Means sangat sensitif terhadap lokasi awal centroid. Dengam kata lain, inisialisasi set representasi cluster C yang berbeda dapat mengakibatkan hasil cluster yang berbeda, bahkan pada set data X yang sama. Inisalisasi yang jelek dapat mengakibatkan hasil cluster yang jelek juga.

Contoh kasus ini diambil dari Jurnal yang berjudul “Implementasi Metode K-Means Clustering Untuk Analisa Prestasi Siswa Berdasarkan Data Siswa di SMA Negeri 1 Grogol” oleh Adhen Bagus Putro Utomo. Dalam contoh kasus ini yaitu untuk analisa prestasi siswa berdasarkan data siswa.

Data yang dianalisa adalah data siswa pada SMA Negeri 1 Grogol. Data yang dikelompokkan dengan menggunakan proses data mining, K-Means Clustering data siswa dari data nilai mata pelajaran Bahasa Indonesia pada tabel 2.3 dan data pengelompokkan siswa pada tabel berikut.

Tabel 2. 3 Data Nilai

No	Nama siswa	Nilai
1	A. Rizki Muzaki	80
2	Bety Chariska A. A.	83
3	Byron Khoirul A.	75
4	Diah Palupi P.	78
5	Diky Octaviani N.	77

Tabel data nilai di atas adalah contoh tabel data nilai matapelajaran Bahasa Indonesia, dari data nilai di atas maka akan dihitung dan diproses dengan metode *K-Means* sebagai berikut :

M1 = 8,0
M2 = 8,3
M3 = 7,5
M4 = 7,8
M5 = 7,7
K = 2
C1 = (4,5) C2 = (7,5)

Keterangan:

M1 adalah data nilai ke 1, M2 adalah data nilai ke 2, dan seterusnya. K adalah jumlah data yang akan di kelompokan, C_1 adalah kelompok data ke 1 dan C_2 adalah kelompok data ke 2. Setelah data nilai didapati kita langsung masuk dalam perhitungan *K-Means* Dengan rumus *Encludian Matrix* antara titik $a = (a_1, a_2, a_3, a_4, \dots, a_n)$ dan titik $b = (b_1, b_2, b_3, b_4, \dots, b_n)$ adalah : $(b_i - a_i)^2$

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$

$$M1\ S_{c1} = \sqrt{(8-4)^2 + (0-5)^2} = \sqrt{(4)^2 + (-5)^2} = \sqrt{16+25} = \sqrt{41}$$

$$M2\ S_{c2} = \sqrt{(8-7)^2 + (3-5)^2} = \sqrt{(1)^2 + (-2)^2} = \sqrt{2+4} = \sqrt{6}$$

$$M3\ S_{c1} = \sqrt{(7-4)^2 + (5-5)^2} = \sqrt{(3)^2 + (0)^2} = \sqrt{6+0} = \sqrt{6}$$

$$M3\ S_{c2} = \sqrt{(7-7)^2 + (5-5)^2} = \sqrt{(0)^2 + (0)^2} = \sqrt{0} = 0$$

$$M4\ S_{c1} = \sqrt{(7-4)^2 + (8-5)^2} = \sqrt{(3)^2 + (3)^2} = \sqrt{9+9} = \sqrt{18}$$

$M4\ S_{c2} = \sqrt{(7-7)^2 + (8-5)^2} = \sqrt{(0)^2 + (3)^2} = \sqrt{0+9} = \sqrt{9}$

$M5\ S_{c1} = \sqrt{(7-4)^2 + (7-5)^2} = \sqrt{(3)^2 + (2)^2} = \sqrt{9+4} = \sqrt{13}$

$M5\ S_{c2} = \sqrt{(7-7)^2 + (7-5)^2} = \sqrt{(0)^2 + (2)^2} = \sqrt{0+4} = \sqrt{4}$

Dari perhitungan data diatas maka dapat disimpulkan bahwa data nilai siswa M1,M2,M3,M4, dan M5 dinyatakan masuk kelompok ke 2 atau Sc_2 karena dari hasil perhitungan diatas nilai terdekat atau terkecil seluruhnya berada dikelompok ke 2 atau Sc_2. Setelah data awal sudah di ketahui kelompoknya maka proses selanjutnya adalah penjumlahan dan pembagian seperti berikut ini:

M1 = 8,0 masuk kelompok ke 2 atau Sc_2
M2 = 8,3 masuk kelompok ke 2 atau Sc_2
M3 = 7,5 masuk kelompok ke 2 atau Sc_2
M4 = 7,8 masuk kelompok ke 2 atau Sc_2
M5 = 7,7 masuk kelompok ke 2 atau Sc_2

Semua data yang masuk kelompok 2 lalu dijumlah, lalu dibagi 5 karena 5 adalah jumlah banyaknya data yang masuk claster ke 2 atau Sc_2, contoh perhitungannya seperti ini.

8,0 + 8,3 + 7,5 + 7,8 + 7,7 = 39,3
39 : 5 = 7,8
3 : 5 = 0,6

Dari perhitungan di atas dapat di ketahui bahwa titik centroid awal telah berubah, dari yang tadinya titik awal centroidnya adalah Sc1 = (4,5) dan Sc2 = (7,5) kini telah berubah menjadi Sc1 = (7,8) dan Sc2 = (0,6) Karena titik awal centroid telah berubah, lakukan perhitungan lagi sampai titik centroidnya tidak berubah atau bisa juga seperti ini. Jika titik centroid awal berubah maka kita akan melakukan perhitungan yang sama dengan perhitungan yang sebelumnya tadi, akan tetapi dalam perhitungan tahap ke 2 ini kita akan menggunakan titik centroid baru yaitu Sc1 = 7,8 dan Sc2 = 0,6 . setelah dalam perhitungan tahap ke 2 dan hasilnya titik centroid tidak berubah maka perhitungannya selesai, dan hasil pengelompokan siswa berprestasi dan kurang berprestasi dapat di ketahui, contoh seperti berikut ini :

M1=8,0 M2=8,3 M3=7,5 M4=7,8 M5=7,7

K=2 Sc1 =7,8 Sc2 =0,6

$M1\ S_{c1} = \sqrt{(8-7)^2 + (0-8)^2} = \sqrt{(1)^2 + (-8)^2} = \sqrt{1+64} = \sqrt{65}$

$M1\ S_{c2} = \sqrt{(8-0)^2 + (0-6)^2} = \sqrt{(8)^2 + (-6)^2} = \sqrt{64+36} = \sqrt{100}$

$M2\ S_{c1} = \sqrt{(8-7)^2 + (3-8)^2} = \sqrt{(1)^2 + (25)^2} = \sqrt{1+25} = \sqrt{26}$

$M2\ S_{c2} = \sqrt{(8-0)^2 + (3-6)^2} = \sqrt{(8)^2 + (-3)^2} = \sqrt{64+9} = \sqrt{73}$

$M3\ S_{c1} = \sqrt{(7-7)^2 + (5-8)^2} = \sqrt{(0)^2 + (-3)^2} = \sqrt{0+9} = \sqrt{9}$

$M3\ S_{c2} = \sqrt{(7-0)^2 + (5-6)^2} = \sqrt{(7)^2 + (-1)^2} = \sqrt{49+1} = \sqrt{50}$

$M4\ S_{c1} = \sqrt{(7-7)^2 + (8-8)^2} = \sqrt{(0)^2 + (0)^2} = \sqrt{0+0} = \sqrt{0}$

$M4\ S_{c2} = \sqrt{(7-0)^2 + (8-6)^2} = \sqrt{(7)^2 + (2)^2} = \sqrt{49+4} = \sqrt{53}$

$M5\ S_{c1} = \sqrt{(7-7)^2 + (7-8)^2} = \sqrt{(0)^2 + (-1)^2} = \sqrt{0+1} = \sqrt{1}$

$M5\ S_{c2} = \sqrt{(7-0)^2 + (7-6)^2} = \sqrt{(7)^2 + (1)^2} = \sqrt{49+1} = \sqrt{50}$

$8,0 + 8,3 + 7,5 + 7,8 + 7,7 = 93,3$

$93 : 5 = 7,8$

$3 : 5 = 0,6$

Setelah dihitung kembali menggunakan titik centroid yang baru dan ternyata hasil klusternya tetap, maka perhitungan selesai.

Tabel 2. 4 Pengelompokkan Siswa

No	Nama Siswa	Nilai	Kelompok

1	A. Rizki Muzaki	80	Berprestasi
2	Bety Chariska A.	83	Berprestasi
3	Byron Khoirul A.	75	Kurang berprestasi
4	Diah Palupi P.	78	Berprestasi
5	Diky Octaviani	77	Kurang berprestasi

Tabel 2.4 pengelompokan siswa di atas adalah tabel hasil dari perhitungan menggunakan metode *K-Means* yang telah di ketahui hasilnya dan telah dikelompokan.

Berdasarkan penelitian di atas, maka dapat membantu dalam pengambilan keputusan untuk analisa prestasi siswa berdasarkan data siswa, dan hasilnya akan dikelompokkan sesuai kelompok yang diinginkan seperti kelompok siswa berprestasi dan kelompok siswa yang kurang berprestasi.

C. Pengelompokan Siswa

Pengelompokkan adalah proses, cara, perbuatan mengelompokkan. Potensi adalah kemampuan yang mempunyai kemungkinan untuk dikembangkan, kekuatan, kesanggupan, daya.

Menurut (Djamarah, 2012) prestasi yaitu hasil dari suatu kegiatan yang telah dikerjakan, diciptakan, yang menyenangkan hati yang diperoleh dengan jalan keuletan kerja, baik secara individual maupun kelompok dalam bidang kegiatan tertentu. Jadi, Pengelompokkan siswa berpotensi adalah suatu proses pengelompokkan dimana siswa yang bepotensi dicalonkan sebagai untuk masuk organisasi.

D. Tinjauan Pustaka

Tabel 2. 5 Tinjauan Pustaka

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
1	Hamdan Yuwafi, Fitri Marisa, Indra Darma Wijaya	Implementasi Data Mining Untuk Menentukan Santri Berprestasi Di Pp.Manaarulhuda	Data nilai santri perlu dikelompokkan untuk mempermudah	Jurnal SPIRIT Vol. 11 No. 1 Mei 2019, hal 22 - 29, Program Studi Teknik	Dari penelitian ini peneliti mengembangkan aplikasi untuk mengelompokkan

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
		Dengan Metode Clustering Algoritma K-Means	dalam pengukuran prestasi dengan jangkauan kelompok nilai tertentu. Hasil pengelompokan nilai ini dapat digunakan untuk membuat kebijkan menentukan santri yang berprestasi.	Informatika , Universitas Widyagama Malang	siswa yang berprestasi yang diterapkan ke dalam metode K-Means. Aplikasi dengan metode Algoritma K-Means terbukti mampu mengolah data masukan berupa nilai Nahwu, Shorof, Ahlak, Hafalan, Kehadiran dan Khidmat menjadi sebuah proses penilaian siswa yang akan dipilih sehingga proses Pengelompokkan siswa berprestasi menjadi lebih cepat dan akurat.
2	Anggoro Eko Wicaksono	IMPLEMENTASI DATA MINING DALAM PENGELOMPOKKAN DATA PESERTA DIDIK DI SEKOLAH UNTUK MEMPREDIKSI CALON PENERIMA BEASISWA DENGAN MENGGUNAKAN	Dengan banyaknya jumlah peserta didik, tentu pemantauan yang dilakukan secara manual tidak akan efektif, sehingga peserta didik yang memiliki nilai	Jurnal Teknologi Rekayasa Volume 21 No.3, Desember 2016, Program Studi Teknik Informatika, Universitas Gunadarma	Dari penelitian ini peneliti mengembangkan aplikasi untuk mengelompokkan peserta didik sekolah agar dapat memprediksi siswa yang berprestasi

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
		ALGORITMA K-MEANS (STUDI KASUS SMAN 16 BEKASI)	akademiknya baik atau yang orangtuanya berpenghasilan kurang dari cukup tidak semuanya terpantau dan sulit diprediksi untuk mendapatkan beasiswa setelah lulus dari sekolah. Oleh karena itu dibutuhkan suatu metode untuk mengelompokkan peserta didik tersebut apakah layak mendapatkan beasiswa berdasarkan nilai akademiknya beasiswa berdasarkan nilai akademiknya atau berdasarkan gaji orangtuanya.		dengan menggunakan Algoritma K-Means. Pengujian aplikasi dengan menggunakan *black boxtesting* dan *User Acceptance Test*.
3	Jaroji, Danuri, Fajri Profesio Putra	K-Means Untuk Menentukan Calon Penerima Beasiswa Bidik Misi Di POLBENG	Penyaluran beasiswa bidik misi kepada mahasiswa yang berasal dari keluarga kurang	JURNAL INOVTEK POLBENG - SERI INFORMATIKA, Vol. 1, No. 1 ,	Dari penelitian ini peneliti mengembangkan aplikasi untuk menentukan calon penerima

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
			mampu harus melalui seleksi yang melibatkan kriteria-kriteria tertentu. Kriteria tersebut seperti penghasilan orang tua, status kepemilikan rumah, kondisi rumah, jumlah tanggungan orang tua, status orang tua dan prestasi akademik.	Juni 2016. Politeknik Negeri Bengkalis	beasiswa bidik misi dengan menggunakan Algoritma K-Means. Implementasi K-Means menggumnakan aplikasi WEKA yang bertujuan untuk membandingkan hasil dengan perhitungan secara teoritis dengan hasil yang didapatkan dengan proses di Weka Interface ini.
4	Green F Mandias, Green A Sandag, Susi Susanti dan Haryanto Reza Musak	Penerapan Algoritma K-Means Untuk Analisis Prestasi Akademik Mahasiswa Fakultas Ilmu Komputer Universitas Klabat	Fakultas Ilmu Komputer adalah salah satu fakultas yang memiliki 408 Mahasiswa yang aktif di Fakultas Ilmu Komputer. Dengan begitu banyak mahasiswa yang ada saat ini, sehingga peneliti memanfaatkan data mahasiswa	Cogito Smart Journal/VOL. 3/NO. 2/DEC 2017. Program Studi Sistem Informasi, Universitas Klabat, Airmadidi	Dari penelitian ini peneliti mengembangkan aplikasi untuk menganalisis prestasi akademik fakultas ilmu komputer di Universitas Klabat dengan Menggunakan metode K-Means.

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
			untuk dapat mengetahui mahasiswa yang memiliki prestasi dimata kuliah yang telah dikontrak dan telah lulus pada matakuliah yang sudah diambil.		
5	Fitri Yunita	Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru (Studi Kasus : Universitas Islam Indragiri)	Proses penerimaan mahasiswa baru Universitas Islam Indragiri menghasilkan data mahasiswa yang sangat berlimpah berupa data profil mahasiswa dan data lainnya. Hal tersebut terjadi secara berulang dan menimbulkan penumpukan terhadap data mahasiswa baru, sehingga mempengaruhi pencarian informasi terhadap data tersebut.	Jurnal SISTEMASI, Volume 7, Nomor 3 September 2018 : 238 – 249. Program Studi Sistem Informasi, Fakultas Teknik dan Ilmu Komputer Universitasi Islam Indragiri (UNISI)	Dari penelitian ini peneliti mengembangkan aplikasi untuk melakukan pengelompokkan terhadap data penerimaan mahasiswa baru dengan Menggunakan metode K-Means. Pengolahan data menggunakan RapidMiner 5.3 yang digunakan untuk membantu menemukan nilai yang akurat.
6	Ai Ilah Warnilah	Analisis Algoritma K-Means Clustering Untuk Pemetaan Prestasi Siswa Studi Kasus SMP Negeri 1 Sukahening	Seiring dengan terus bertambahnya jumlah data siswa setiap tahun, maka jumlah data yang siswa yang terus meningkat sehingga penumpukan data yang belum diolah dengan optimal untuk menggali	Indonesian Journal on Computer and Information Technology Vol 1 No 1 Mei 2016	Pengolahan data melalui metode Algoritma K-Means untuk menentukan jarak terdekat menggunakan Euclidian Distance lebih optimal dibandingkan dengan menggunakan Mahattan

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
			informasi dan pengetahuan baru melalui pola - pola yang terbentuk dari penumpukan data tersebut.		Distance dan Chbychep Distance dalam mengelompokkan prestasi siswa.
7	Fitri Larasati Sibuea dan Andy Sapta	Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustering	Seiring dengan terus bertambahnya jumlah data siswa yang berprestasi setiap tahun, maka tidak dapat diidentifikasi pengelompokkan prestasi siswa yang tinggi, menengah, dan cukup. Hal itu menyebabkan kurangnya informasi mengenai siswa - siswa yang berprestasi terutama pada proses pemetaannya.	JURTEKSI (Jurnal Teknologi dan Sistem Informasi) Vol. IV No. 1, Des 2017, hlm. 85 – 92 Program Studi Sistem Informasi, STMIK Royal.	Dari penelitian ini peneliti mengembangkan aplikasi untuk melakukan pemetaan siswa berprestasi dengan menggunakan metode K-Means. Pengolahan data siswa berprestasi dengan menggunakan software *Rapid Miner*, dengan adanya software Rapid Miner dalam penelitian ini maka keakuratan data akan cukup baik terhadap permasalahan yang terjadi terkait dengan prestasi siswa dengan hasil 70% dapat mengenali data pada 10 data yang digunakan sebagai sampel.
8	Nurul Rohmawati W, Sofi Defiyanti, Mohamad Jajuli	Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa	Tingginya biaya perkuliahan mempengaruhi kelangsungan kegiatan belajarnya mahasiswa di sebuah instansi pendidikan tinggi. Hal ini meyebabkan banyaknya mahasiswa yang mengajukan cuti	Jurnal Ilmiah Teknologi Informasi Terapan Volume I, No 2, 30 April 2015. Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang.	Dari penelitian ini peneliti mengembangkan aplikasi untuk mengklasifikasi mahasiswa yang berhak menerima beasiswa, mahasiswa yang di pertimbangkan menerima dan mahasiswa yang tidak berhak menerima

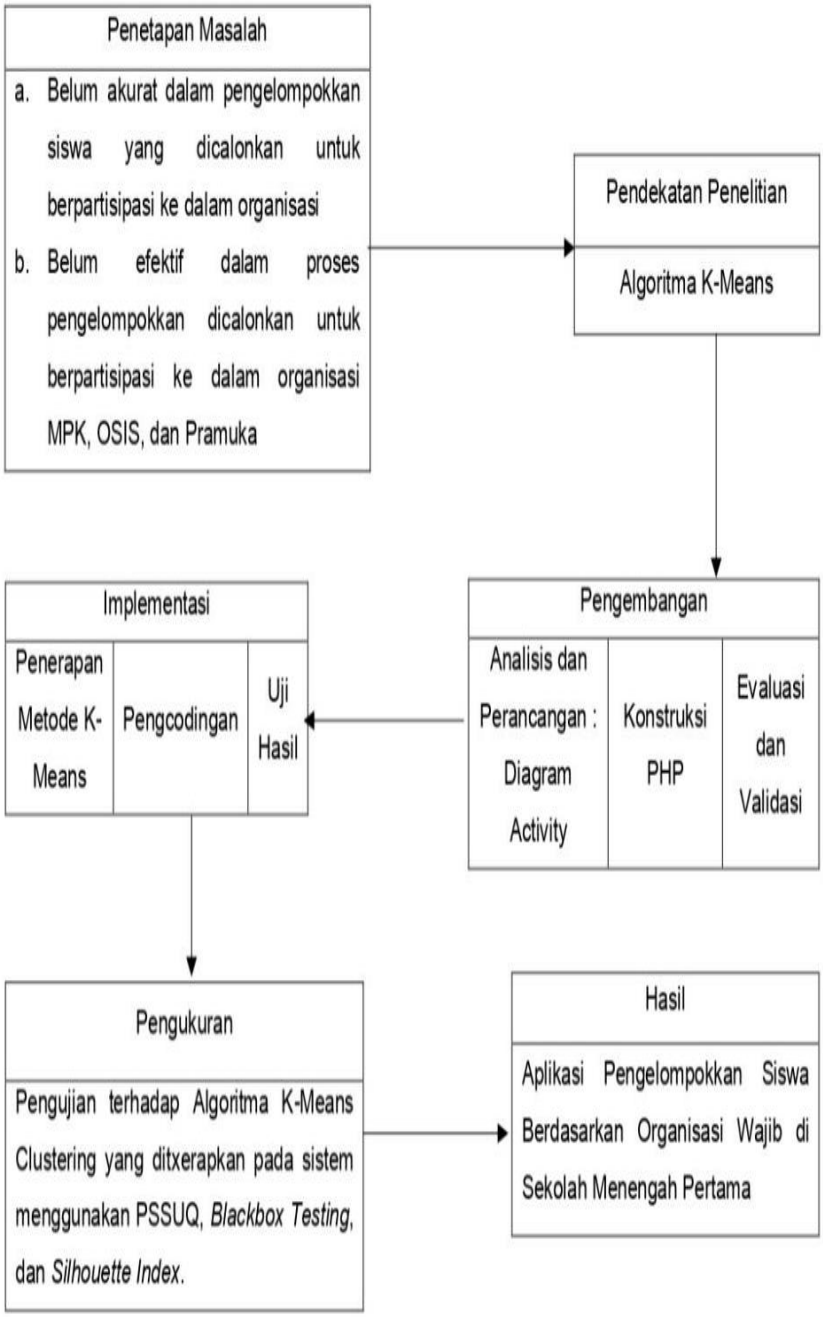
No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
			akademik dan bahkan *dropout*.		beasiswa dengan metode K-Means. Pengolahan data untuk mengklasifikasi pelamar beasiswa dengan menggunakan *Rapid Miner Studio 5.*dengan menggunakan aplikasi ini dapat diketahui Diketahui nilai purity pada dataset data kodifikasi sebagian untuk hasil cluster algoritma k-means sebesar 0.611 atau 61.11%. Pada dataset kodifikasi keseluruhan nilai purity hasil cluster algoritma k-means sebesar 0.806 atau sebesar 80.56%. Untuk dataset data asli nilai purity hasil cluster algoritma k-means sebesar 0.750 atau 75%. Maka dapat disimpulkan bahwa tingkat akurasi clustering hasil cluster algoritma k-means berdasarkan nilai purity measure, dataset yang dikodifikasi keseluruhan lebih baik dari pada dataset yang di kodifikasi sebagian dan dataset data asli.

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
9	Risa Helilintar, Intan Nur Farida	Penerapan Algoritma K-Means Clustering Untuk Prediksi Prestasi Nilai Akademik Mahasiswa	UNP (Universitas Nusantara PGRI) Kediri khususnya Fakultas Teknik Program Studi Teknik Informatika Mahasiswanya semakin meningkat setiap tahunnya. Semakin meningkatnya jumlah mahasiswa yang diluluskan setiap tahunnya menyebabkan banyaknya data mahasiswa yang perlu diolah sehingga pihak prodi menyebabkan kesulitan dalam mengolahnya dan megelompokkan data tersebut.	Jurnal Sains dan Informatika Volume 4, Nomor 2, November 2018. Fakultas Teknik, Prodi Teknik Informatika, Universitas Nusantara PGRI Kediri.	Dari penelitian ini peneliti mengembangkan aplikasi untuk mengklasifikasi data mahasiswa yang perlu diolah, sehingga diperoleh hasil klasifikasi prestasi nilai akademik mahasiswa dari jumlah mahasiswa yang diluluskan setiap tahunnya. Pengolahan data menggunakan RapidMiner dan hasilnya sama dengan perhitungan Analisa Algoritma K-Means yang dilakukan, dan hasilnya cukup efisien dan efektif.
10	Sessy Rewetty Revilla, Analisa Fitria Azhar, Lathifaturrahman	Penggunaan Metode K-Means Clustering Dalam Seleksi Mahasiswa Calon Penerima Beasiswa Berprestasidi Iain Antasari Banjarmasin (Studi Kasus Jurusan Pendidikan Matematika)	Proses penyeleksian beasiswa oleh pihak jurusan masih manual yaitu dengan cara menurutkan peringkat berdasarkan IPK tertinggi	Tashwir Vol. 3 No. 7, Juli – September 2015. Fakultas Tarbiyah dan Keguruan IAIN Antasari	Dari penelitian ini peneliti mengembangkan aplikasi untuk mengklasifikasi data mahasiswa yang perlu diolah, sehingga diperoleh hasil klasifikasi prestasi. Penelitian dengan menggunakan metode Algoritma K-Means ini peneliti memberikan alternatif lain yaitu variabel IPK Mata Kuliah Keahliannya (Jurusannya), kemudian IPK

No	Nama Peneliti	Judul Penelitian	Permasalahan	Jurnal	Kontribusi
					dari Mata Kuliah Dasar Keahlian (ketarbiyahan), selanjutnya Mata Kuliah Umum (Institut) sehingga proses Pengelompokkan calon penerima beasiswa berprestasi menjadi lebih cepat dan akurat.

E. Kerangka Pemikiran

Berdasarkan dukungan landasan teoritis yang diperoleh dari eksplorasi teori yang dijadikan rujukan penelitian, maka dapat disusun kerangka pemikiran pada tabel 2.8 berikut :



Tabel 2. 6 Kerangka Pemikiran

Keterangan kerangka pemikiran pada gambar 2.6 dapat dijelaskan sebagai berikut:

1. Penetapan masalah mencakup fenomena yaitu pihak manajemen pesantren masih sulit dalam mengelompokkan siswa yang akan dicalonkan untuk berpartisipasi ke dalam organisasi
2. Pendekatan penelitian yaitu metode yang diterapkan atau digunakan yaitu metode K-Means
3. Pengembangan yaitu tahap analisis dan perancangan model untuk activity diagram, kemudian menentukan bahasa pemrograman dengan konstruksi PHP, dan kemudian melakuan evaluasi serta validasi
4. Implementasi yaitu tahap menerapkan metode K-Means ke dalam sistem aplikasi dan pembuatan coding, kemudian melakukan uji hasil dari sistem informasi tersebut
5. Pengukuran yaitu melakukan pengujian terhadap metode K-Means yang diterapkan ke sistem menggunakan kuesioner PSSUQ untuk uji pengguna, *Blackbox Testing* uji ahli dan *Silhouette Index* untuk evaluasi cluster.
6. Hasil yaitu sistem informasi yang menampilkan hasil pengelompokkan siswa berdasarkan organisasi wajib di Sekolah Menengah Pertama

F. Hipotesis Penelitian

Hipotesis dalam penelitian ini penerapan metode Algoritma K-Means Clustering diduga dapat mengelompokkan siswa yang dicalonkan untuk berpartisipasi ke dalam organisasi.