

## BAB II KERANGKA TEORITIS

### A. Tinjauan Pustaka

Penelitian rujukan merupakan acuan yang dibutuhkan seorang peneliti untuk melakukan penelitian. Penelitian rujukan pada penelitian ini diambil berdasarkan kesamaan metode yaitu dengan naive bayes, banyak penelitian yang menggunakan metode ini dalam berbagai kasus. Antara lain adalah :

1. Penelitian dilakukan oleh **Asrul Sani** yang berjudul **“PENERAPAN METODE K-MEANS CLUSTERING PADA PERUSAHAAN”** tahun 2018. Sumber Data yang digunakan dalam penelitian adalah data primer dan data sekunder. Sumber data primer merupakan sumber data yang diperoleh secara langsung dari sumber asli dan tidak melalui media perantara. Data historis transaksi penjualan yang digunakan diperoleh secara langsung dari objek penelitian melalui wawancara dan dokumentasi. Sedangkan data sekunder merupakan sumber data penelitian yang diperoleh secara tidak langsung melalui media perantara diperoleh dan dicatat oleh pihak lain. Data sekunder pada umumnya berupa bukti catatan atau laporan historis yang dipublikasikan. Data sekunder yang dimaksud dalam penelitian ini adalah sumber data yang digunakan untuk menunjang kelengkapan teori data primer. Populasi penelitian ini adalah data historis transaksi penjualan bulan Januari s/d Desember 2017 dengan data master kode koleksi/artikel yang mengandung kategori Jas dan Celananya saja yang memiliki 37 jenis produk dan berjumlah 468 transaksi. Sampel adalah sebagian jumlah obyek yang diteliti. Nilai *centroid* ditentukan dengan membagi menjadi 4 kelompok dimana C1 merupakan level transaksi tertinggi, C2 merupakan level transaksi terendah, C3 dan C4 merupakan level transaksi rata rata 1 dan rata rata 2. Dari hasil *Clustering* didapatkan untuk nilai C1 berjumlah 10 kode artikel, nilai untuk C2 berjumlah 4 kode artikel, nilai untuk C3 berjumlah 15 artikel dan untuk C4 berjumlah 8 artikel..
2. Penelitian yang dilakukan oleh **Penda Sudarto Hasugian** tahun 2017 yang berjudul **“ANALISA DAN IMPLEMENTASI METODE K-MEANS CLUSTERING DALAM PREDIKSI PERSEDIAAN ALAT KONTRASEPSI (STUDI KASUS : KABUPATEN DELISERDANG)”**. Data persediaan alat kontrasepsi pada Badan Keluarga Berencana Dan Pemberdayaan Perempuan Kab. Deli serdang selama ini hanya berfungsi sebagai arsip bagi Badan KB, sehingga data persediaan alat kontrasepsi tersebut tidak dapat dimanfaatkan seperti untuk mendukung pengembangan pengadaan persediaan alat kontrasepsi pada Badan KB dan Pemberdayaan Perempuan

Kab. Deli Serdang. Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan dalam database. Data Mining adalah proses ekstraksi informasi dari kumpulan data melalui penggunaan algoritma dan teknik yang melibatkan bidang ilmu statistik, mesin pembelajaran, matematika, dan kecerdasan buatan. Data mining digunakan untuk ekstraksi informasi penting yang tersembunyi dari dataset yang besar. Dengan adanya data mining maka akan didapatkan suatu permata berupa pengetahuan di dalam kumpulan data-data yang banyak jumlahnya. Kmeans *Clustering* merupakan salah satu metode data *Clustering* non-hirarki yang mengelompokkan data dalam bentuk satu atau lebih *cluster*/kelompok. Menggunakan pendekatan *partitional Clustering*. Tiap *cluster* dihubungkan dengan sebuah *centroid* (titik pusat). Tiap titik ditempatkan ke dalam *cluster* dengan *centroid* terdekat. Jumlah *cluster*,  $k$ , harus ditentukan. Penerapan metode K-Means *Clustering* untuk mengetahui persediaan alat kontrasepsi yang paling banyak masuk dan keluar pada Badan Keluarga Berencana dan Pemberdayaan Perempuan Kab. Deli Serdang. Algoritma KMeans *Clustering* pada rapidminer dimulai dengan penginputan data persediaan alat kontrasepsi menjadi database pada Ms.Excel dengan ekstensi csv. Mengelompokkan data dengan algoritma k-means *Clustering*, dilakukan dengan cara menentukan jumlah *cluster*, hitung jarak terdekat dengan pusat *cluster*. Data dengan jarak terdekat menyatakan anggota *cluster* tersebut, dilakukan perhitungan kembali sampai data tidak berpindah pada *cluster* lain, untuk meminimalkan fungsi objektif..

3. Penelitian yang dilakukan oleh **TEGUH WIBOWO** tahun 2018 yang berjudul "**Penerapan Data Mining Pemilihan Siswa Kelas Unggulan dengan Metode K-Means *Clustering* di SMP N 02 Tasikmadu**". SMP N 02 Tasikmadu merupakan sekolah yang berada di kota Karanganyar. Sekolah ini memiliki sistem pembelajaran yang baik yaitu program kelas siswa unggulan. SMP N 02 Tasikmadu mengalami peningkatan penerimaan pendaftar siswa baru dalam setiap tahunnya. Banyaknya siswa yang mendaftar maka dalam pemilihan siswa kelas unggulanpun mengakibatkan kesulitan dalam menentukan kelas unggulan yang sesuai dengan bakat kemampuan siswa. Oleh sebab itu penerapan data mining ini dilakukan untuk membantu keputusan dalam memilih pengelompokkan kelas unggulan dengan acuan nilai-nilai yang dimiliki siswa dengan menggunakan metode k-means *Clustering*. Sebagai penerapan metode *Clustering* untuk data perhitungan algoritma k-means yang digunakan adalah nilai raport siswa dan

masing-masing data nilai digunakan sebagai atribut. Atribut-atribut yang dipilih diterapkan menggunakan metode k-means *Clustering* untuk menghasilkan 5 *cluster* yang diambil 3 *cluster* untuk kelas unggulan. Hasil dari penelitian ini menunjukkan bahwa algoritma k-means mampu menghasilkan pemilihan dan pembagian kelas unggulan sesuai nilai kemampuan siswa.

4. Penelitian yang dilakukan oleh **Teguh Nurhadi Suharsono, Maulana Akbar** yang berjudul **“PREDIKSI MINAT KONSUMEN SESUAI MUSIM MENGGUNAKAN ALGORITMA K-MEANS PADA PERANGKAT LUNAK PEMESANAN TEMPAT DAN MAKANAN ONLINE”** tahun 2017. Banyak menu makanan yang tidak terjual menyebabkan restoran merugi dikarenakan bahan makanan yang ada jadi tidak terpakai. Maka dalam hal ini K-Means dapat digunakan untuk meng-*cluster* minat konsumen dimana minat tersebut akan disesuaikan dengan musim. Algoritma K-Means akan digunakan pada perangkat lunak untuk menemukan minat konsumen dari data penjualan yang ada, dalam hal ini bisa dikatakan sebagai data mining dimana data tersebut akan dibuat dan diurutkan berdasarkan minat pelanggan seperti musim panas ataupun musim hujan. Dan penentuan permasalahan tadi berdasarkan musim dikarenakan minat konsumen akan lebih mudah dikalkulasikan mengikuti musim yang ada di Indonesia dimana Indonesia sendiri adalah negara beriklim tropis yang dimana memiliki menu makanan yang beragam. Dengan adanya pengelompokan data dari data penjualan diharapkan rumah makan atau restoran dapat menentukan menu makanan dan minuman yang sesuai dengan minat konsumen dimana menu makanan tersebut akan disesuaikan dengan musim yang ada di Indonesia yaitu musim panas dan musim hujan. Pengimplementasian algoritma K-Means *Clustering* pada menu makanan berdasarkan data penjualan dapat dilakukan dan hasil yang didapat sesuai dengan apa yang diharapkan yang berupa laporan untuk menjadi bahan pertimbangan bagi manajemen tingkat atas untuk menentukan menu makanan sesuai dengan musim. Penerapan algoritma K-Means pada Perangkat lunak Pemesanan dan Tempat Makan Online telah sesuai dengan kriteria yang dimiliki K-Means *Clustering*, dimana data menu makanan yang diminati dan menu makanan yang tidak diminati bisa ditemukan oleh K-Means dan hal tersebut dapat membantu pihak restoran agar tidak mengalami kerugian yang dikarenakan adanya bahan makanan yang tidak terpakai. Hasil dari perhitungan algoritma K-Means dapat dijadikan sebagai produk yang disarankan karena data yang diambil didapatkan dari *cluster* pertama yaitu *cluster* yang kriterianya adalah produk yang banyak terjual atau dalam hal ini

adalah produk yang banyak dibeli oleh konsumen. Hal tersebut pula berhasil diimplementasikan, serta hasil K-Means tersebut juga dapat menjadi daya jual untuk memperlihatkan atau mengenalkan menu makanan terbaik yang dimiliki oleh restoran. Dalam hal ini dapat disimpulkan data yang didapat dari hasil perhitungan Kmeans dapat dijadikan menu rekomendasi untuk pelanggan agar pelanggan dapat mengetahui menu rekomendasi dari pihak restoran..

5. **Elly Muningsih dan Sri Kiswati** telah melakukan penelitian yang berjudul **“Penerapan Metode K-Means Untuk *Clustering* Produk Online Shop Dalam Penentuan Stok Barang”** tahun 2015. Manajemen stok yang dilakukan secara tidak akurat dan asal-asalan akan menyebabkan biaya simpan yang tinggi dan tidak ekonomis, karena bisa terjadi kekosongan atau kelebihan produk tertentu. Hal ini tentu akan sangat merugikan semua pelaku usaha seperti halnya online shop. Penelitian ini bertujuan untuk mengelompokkan produk yang dijual pada online shop Ragam Jogja menjadi beberapa *cluster* untuk mengetahui produk mana yang paling diminati sehingga jumlah stok harus banyak, produk diminati untuk jumlah stok sedang dan produk kurang diminati untuk jumlah stok sedikit. Populasi target dalam penelitian ini adalah data historis transaksi penjualan bulan Januari 2011 s/d Mei 2012 dengan data master kode produk yang hanya mengandung kategori Batik saja yang memiliki 31 jenis produk dan berjumlah 235 transaksi. Metode yang digunakan pada penelitian ini adalah metode K-Means yang merupakan salah satu metode terbaik dan paling populer dalam algoritma *Clustering* dimana K-Means mencari partisi yang optimal dari data dengan meminimalkan kriteria jumlah kesalahan kuadrat dengan prosedur iterasi yang optimal. Variabel yang digunakan adalah kode produk, jumlah transaksi, volume penjualan dan rata-rata penjualan. Penelitian juga didukung dengan software Rapidminer untuk pengolahan data dengan metode K-Means. Hasil akhir dari penelitian adalah berupa suatu program aplikasi yang dapat mengelompokkan produk menjadi kategori jumlah stok banyak, sedang dan sedikit berdasarkan transaksi penjualan. Berdasarkan pengolahan data yang dilakukan dengan atribut kode produk, jumlah transaksi, volume penjualan dan rata-rata penjualan, dihasilkan 3 kelompok produk yang paling diminati berjumlah 3 produk untuk jumlah stok banyak, 11 produk diminati untuk jumlah stok sedang dan 17 produk kurang diminati untuk jumlah stok sedikit.
6. **Silvi Agustina, Dhimas Yhudo, Hadi Santoso, Nofiadi Marnasusanto, Arif Tirtana, Fakhris Khusnu** melakukan penelitian dengan judul **“*CLUSTERING* KUALITAS BERAS BERDASARKAN CIRI FISIK MENGGUNAKAN**

**METODE K-MEANS**” Saat ini pemeriksaan kualitas beras telah dilakukan secara manual oleh inspektur yang telah berpengalaman. Dengan cara ini memiliki kelemahan seperti: adanya subjektivitas penilaian mutu antara pengamat yang satu dengan yang lain.adanya kelelahan fisik jika pengamat bekerja terlalu lama, sehingga menyebabkan hasil pengamatan tidak konsisten, dan waktu yang dibutuhkan untuk pengamatan mutu lebih lama. Sehubungan dengan permasalahan diatas, maka diperlukan cara untuk menentukan klasifikasi mutu beras dengan cepat, akurat dan mudah untuk dioperasikan, sehingga dapat meningkatkan efisiensi pengkelasan mutu fisik beras. K-Means merupakan metode dengan tipe klasifikasi unsupervisi dimana mengelompokkan data menjadi satu atau lebih klaster. K-Means memodelkan dataset menjadi klaster-klaster dimana data pada satu klaster memiliki karakteristik yang sama dan memiliki karakteristik yang berbeda dari klaster lain. Penelitian ini bertujuan untuk membuat sistem penunjang keputusan untuk menentukan klasifikasi mutu beras. Data sampel beras yang digunakan pada penelitian ini sebanyak 20 dengan fitur  $X_1$  yaitu panjang beras (mm) dan  $X_2$  yaitu lebar beras (mm). Perhitungan pada program Klasterisasi Kualitas Beras dengan KMeans diawali dengan menentukan jumlah *cluster*. Penelitian ini menggunakan 3 *cluster* untuk menentukan kualitas beras, dengan keterangan bahwa *cluster* 1 adalah beras kualitas buruk, *cluster* 2 adalah beras kualitas sedang, dan *cluster* 3 adalah beras kualitas baik. Setelah menentukan jumlah *cluster*, kemudian menentukan 3 pusat *cluster* awal, yaitu pusat *cluster* 1 (5,5 ; 2), pusat *cluster* 2 (6,2 ; 2,45), dan pusat *cluster* 3 (7 ; 3,2). Langkah selanjutnya adalah mengalokasikan data ke dalam *cluster*, kemudian menghitung jarak setiap data terhadap setiap pusat *cluster* dengan menggunakan Persamaan 1. Suatu data akan menjadi anggota dari suatu *cluster* yang memiliki jarak terkecil dari pusat *cluster*-nya. Setelah itu menghitung pusat *cluster* baru sebagai acuan untuk iterasi berikutnya. Pada iterasi 1 dihasilkan 3 pusat *cluster* baru, pusat *cluster* 1 (5,89333;2,05), pusat *cluster* 2 (6,28199;2,546), dan pusat *cluster* 3 (6,96583;2,99917). Perhitungan pada iterasi selanjutnya sama dengan iterasi 1, hingga posisi data tidak mengalami perubahan. Pusat *cluster* yang dihasilkan pada iterasi 2 yaitu pusat *cluster* 1 (5,74499;2,07), pusat *cluster* 2 (6,26667;2,45667), dan pusat *cluster* 3 (6,96583;2,99917). Pusat *cluster* yang dihasilkan pada iterasi 3 yaitu pusat *cluster* 1 (6,28199;2,546), pusat *cluster* 2 (5,89333;2,05), dan pusat *cluster* 3 (6,96583;2,99917). Pusat *cluster* yang dihasilkan pada iterasi 4 yaitu pusat *cluster* 1 (5,89333;2,05), pusat *cluster* 2 (6,28199;2,546), dan pusat *cluster* 3

(6,96583;2,99917). karena pada iterasi selanjutnya posisi data tidak berubah, maka iterasi dihentikan dan hasil akhir yang diperoleh adalah 3 pusat *cluster* pada iterasi ke-4. K-Means *Clustering* merupakan metode klasterisasi berdasarkan persamaan karakteristik, dan merupakan metode yang sangat berguna karena mampu mentranslasi ukuran persamaan yang intuitif menjadi ukuran yang kuantitatif. Penelitian ini menggunakan 20 data uji, dimana ke-20 data tersebut dibagi menjadi 3 *cluster* dengan keterangan *Cluster* 1 merupakan beras kualitas buruk, *Cluster* 2 beras kualitas sedang, dan *Cluster* 3 beras kualitas baik. Dari hasil penelitian, didapatkan 3 pusat *cluster* akhir yaitu pusat *cluster* 1 (5,89333;2,05), pusat *cluster* 2 (6,28199;2,546), dan pusat *cluster* 3 (6,96583;2,999167) serta dihasilkan validasi sebesar 92,82% yang menunjukkan bahwa program ini dapat dijadikan sebagai acuan dalam klasterisasi kualitas beras..

7. **Ade Bastian, Harun Sujadi, dan Gigin Febrianto** melakukan penelitian dengan judul **“PENERAPAN ALGORITMA K-MEANS CLUSTERING ANALYSIS PADA PENYAKIT MENULAR MANUSIA (STUDI KASUS KABUPATEN MAJALENGKA)”** Kesehatan merupakan hal yang berharga bagi manusia karena siapa saja dapat mengalami gangguan kesehatan, begitu pula pada manusia yang sangat rentan terhadap berbagai macam penyakit namun penyebabnya tidak kita sadari. Algoritma K-means tidak terpengaruh terhadap urutan objek yang digunakan, hal ini dibuktikan ketika penulis mencoba menentukan secara acak titik awal pusat *cluster* dari salah satu objek pada permulaan perhitungan. Jumlah keanggotaan *cluster* yang dihasilkan berjumlah sama ketika menggunakan objek yang lain sebagai titik awal pusat *cluster* tersebut. Namun, hal ini hanya berpengaruh pada jumlah iterasi yang dilakukan. Pengelompokan objek (objek *Clustering*) adalah salah satu proses dari objek mining yang bertujuan untuk mempartisi objek yang ada ke dalam satu atau lebih *cluster* objek berdasarkan karakteristiknya. Penelitian ini mengkaji bagaimana penggunaan Algoritma K-means *Cluster Analysis* dalam studi kasus penyakit menular manusia pada suatu objek. Penelitian ini mengkaji metode K-means *Cluster Analysis* dalam penyakit menular pada manusia berdasarkan set variabel yang dibentuk per kecamatan tiap Puskesmas yang jumlahnya ada 32 Kantor Puskesmas di Kabupaten Majalengka. Dari hasil penelitian yang telah dilakukan dapat disimpulkan bahwa K-means merupakan salah satu metode data *Clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster* atau kelompok. Kelebihan Algoritma K-means diantaranya adalah

mampu mengelompokkan objek besar dan pencilaan obyek dengan sangat cepat sehingga mempercepat proses pengelompokan. Kekurangan Algoritma K-means yaitu. sangat sensitive pada pembangkitan titik pusat awal secara random, hasil pengelompokan bersifat tidak unik (selalu berubah-ubah) dan proses pengerjaannya cepat tetapi keakuratannya tidak dijamin. Dari hasil penelitian dapat ditarik kesimpulan bahwa hasil dari metode Algoritma K-means *Clustering* data mining dapat digunakan untuk metode pengendalian persediaan pada Puskesmas Pandanaran, sehingga apabila akan dilakukan pengadaan persediaan obat pada tahun 2014, petugas dapat melihat daftar Puskesmas terbanyak yang menderita penyakit menular. .

8. **M.Hasyim Siregar** melakukan penelitian dengan judul **“KLASTERISASI PENJUALAN ALAT-ALAT BANGUNAN MENGGUNAKAN METODE K-MEANS (STUDI KASUS DI TOKO ADI BANGUNAN)”** tahun 2018. Dalam persaingan dunia bisnis saat ini, kita dituntut untuk senantiasa mengembangkan bisnis agar selalu bertahan dalam persaingan. Untuk mencapai hal tersebut, ada beberapa hal yang bisa dilakukan yaitu dengan meningkatkan kualitas produk, penambahan jenis produk, dan pengurangan biaya operasional perusahaan dengan cara menggunakan analisis data perusahaan. Data mining adalah sebuah teknologi yang mengotomatisasi proses untuk menemukan pola menarik dan sensitif dari kumpulan-kumpulan data yang besar. Ini memungkinkan pemahaman manusia tentang menemukan pola dan skalabilitas teknik. Toko Adi Bangunan merupakan sebuah toko yang bergerak dalam bidang penjualan bahan-bahan bangunan dan peralatan rumah tangga yang memiliki sistem seperti pada swalayan yaitu pembeli mengambil sendiri barang yang akan dibeli. Dataset yang digunakan dalam penelitian ini adalah hasil dari database pada data penjualan 1 Juli 2017 sampai dengan 9 Agustus 2017 yang berjumlah 10 buah data dengan instrumen berdasarkan stok awal, penjualan dan stok akhir. Berdasarkan eksperimen perhitungan algoritma k-means hasil yang diperoleh sebagai berikut, *Cluster 1=2* anggota *Cluster 2=8* anggota. Berdasarkan hasil *cluster*, maka dilakukan analisa dalam pengelompokan barang. Analisa dilakukan dengan menghitung rata-rata dari setiap anggota *cluster* berdasarkan kategori yang telah ditentukan. Dalam penelitian ini ini dapat disimpulkan beberapa hal yang diperlukan untuk menganalisis penjualan yang terjadi di Toko Adi Bangunan pada 1 Juli 2017 sampai dengan 9 Agustus 2017. Data tersebut diolah menggunakan k-mean *Clustering* untuk menentukan kelompok barang yang laris dan yang kurang laris dalam penelitian tersebut. kesimpulan dari

penelitian ini adalah Penerapan metode *Clustering* dapat menentukan pembelian stok barang yang diperlukan dengan cepat, dari penelitian yang dilakukan diketahui kelompok barang yang laris sebanyak 10 item sehingga prioritas pembelian stok barang diarahkan pada 10 item tersebut, Penerapan metode k-mean dengan menggunakan bantuan software rapid miner menghasilkan item barang yang laris dan yang kurang laris sehingga pihak manajemen dapat menentukan prioritas pembelian stok barang pada barang yang laris.

9. **M. Nanda Variestha Waworuntu, Muhammad Faisal Amin** melakukan penelitian dengan judul **“PENERAPAN METODE K-MEANS PEMETAAN CALON PENERIMA JAMKESDA”**. tahun 2018. Pada Kelurahan Kemuning salah satunya Seksi Kesejahteraan Sosial (KESSOS) terdapat pelayanan masyarakat miskin untuk menerima bantuan Jaminan Kesehatan Daerah (JAMKESDA). Selama ini bagian Seksi KESSOS pada Kelurahan Kemuning belum ada metode yang dapat mengelompokkan tingkat kemiskinan agar penerima bantuan tepat sasaran, sehingga pihak Kelurahan tidak dapat mencegah ketidaktepat sasaran tersebut. Oleh sebab itu, pengelompokan kemiskinan dapat membantu pihak Kelurahan dalam mengambil keputusan yang tepat untuk mencegah ketidaktepat sasaran penerima JAMKESDA. Pada penelitian ini, penerapan metode K-Means diimplementasikan pada aplikasi yang dibuat dengan 2 klaster. Penelitian ini menggunakan sebanyak 440 sampel data. Dari hasil perhitungan Davies Bouldin Index diperoleh nilai penentuan jumlah *cluster* dengan nilai 2 klaster (0.243), 3 klaster (0.256), 4 klaster (0.275). Nilai yang digunakan adalah 2 klaster karena nilai tersebut mendekati 0. *Cluster Model* yang dihasilkan dari pengelompokan data penduduk dengan algoritma K-Means tampak bahwa jumlah *cluster* yang digunakan sebanyak 2 dengan keanggotaan *Cluster 0* terdiri dari 334 item, *Cluster 1* terdiri dari 106 item dengan jumlah keseluruhan adalah sebanyak 440 item. Pada Tabel 2 tampak bahwa jumlah *cluster* yang digunakan sebanyak 3 dengan keanggotaan *Cluster 0* terdiri dari 43 item, *Cluster 1* terdiri dari 87 item, dan *Cluster 2* terdiri dari 310 item dengan jumlah keseluruhan adalah sebanyak 440 item. Pada Tabel 3 tampak bahwa jumlah *cluster* yang digunakan sebanyak 4 dengan keanggotaan *Cluster 0* terdiri dari 60 item, *Cluster 1* terdiri dari 52 item, *Cluster 2* terdiri dari 302, dan *Cluster 3* terdiri dari 26 item dengan jumlah keseluruhan adalah sebanyak 440 item. Percobaan pertama dengan 2 *cluster* yang lebih optimal dikarenakan nilai Performance Vektor (Davies Bouldin Index) mendekati nilai 0 dengan nilai 0.243. n

10. **Yusuf Ramadhan Nasution, Muhammad Eka** melakukan penelitian dengan judul “**PENERAPAN ALGORITMA K-MEANS CLUSTERING PADA APLIKASI MENENTUKAN BERAT BADAN IDEAL**” tahun 2018. Salah satu penerapan teknologi komputer dalam dunia kedokteran adalah untuk menentukan berat badan ideal (BMI) seorang pasien dengan membandingkan berat badan, tinggi badan, serta ukuran kerangka pasien itu sendiri, sehingga dokter bisa menentukan menu diet yang paling cocok untuk pasien tersebut. Hal ini menjadi hal yang sangat penting terutama di bidang kedokteran, seperti bidang kecantikan, atlet, atau bidang yang lain yang menuntut bentuk tubuh ideal seperti model, artis dan lain sebagainya. Sistem penentuan berat badan ideal ini dilakukan dengan perbandingan tinggi badan dengan berat badan, serta ukuran kerangka pasien. Sedangkan algoritma *K-Means Clustering* merupakan sebuah algoritma yang dapat mengelompokkan data berdasarkan nilai-nilai patokan yang diberikan dan mengkalkulasikan kelompok data tersebut. Dengan menggunakan metode *K-Means Clustering*, dapat dilakukan pengelompokan data berat badan ideal dengan cara membandingkan titik pusat masing-masing data dengan titik pusat *cluster* yang telah ditetapkan. Berdasarkan nilai BMI dan Ukuran Kerangka pasien. Nilai BMI digunakan sebagai koefisien X dan nilai Ukuran Kerangka pasien digunakan sebagai koefisien Y serta titik pusat data yang akan dikelompokkan terlebih dahulu.

Persamaan dengan 10 penelitian yang diatas adalah menggunakan metode yang sama yaitu *K-Means*. Perbedaan penelitian ini adalah jumlah *cluster*. *Cluster* dibagi berdasarkan tingkat penjualan sepatunya yaitu tinggi, sedang, rendah. Setelah itu akan direkomendasikan sepatu yang diminati dari 10 merk sepatu tersebut. Dan perbedaan lainnya adalah jenis variabel yang akan digunakan, jumlah variabel yang digunakan adalah 2 yaitu sisa stok sepatu dan jumlah penjualan sepatu. nanti juga akan ada uji kualitas *cluster* dan juga uji ketepatan *cluster*, dan nantinya akan ada rekomendasi merk sepatu yang diminati oleh pelanggan dalam bentuk aplikasi.

## **B. Landasan Teori**

Dalam rangka memperoleh suatu pedoman guna lebih memperdalam masalah, maka perlu dikemukakan suatu landasan teori yang bersifat ilmiah. Dalam landasan teori ini dikemukakan teori yang ada hubungannya dengan materi-materi yang digunakan untuk memecahkan masalah pada penelitian ini.

## 1. Data Mining

(Prasetyo Eko, 2014, p.7), mengatakan bahwa “Data Mining adalah suatu proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan”. (Hermawati, F. A., 2013, pp.1-5), mengatakan bahwa “Data Mining adalah proses yang memperkerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis. Defenisi lain di antaranya adalah pembelajaran berbasis induksi (*induction-based learning*) adalah proses pembentukan defenisi-defenisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik dari konsep-konsep yang akan dipelajari”. Data mining berisi pencarian trend atau pola yang diinginkan dalam *database* yang besar untuk membantu pengambilan keputusan di waktu yang akan datang. Pola-pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisa dsata yang berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mungkin saja menggunakan perangkat pendukung keputusan yang lainnya.

Operasi data mining menurut sifatnya dibedakan menjadi dua, yaitu bersifat prediksi (*prediction driven*) untuk menjawab pertanyaan apa dan sesuatu yang bersifat remang-remang atau transparan. Operasi prediksi digunakan untuk validasi hipotesis, *querying* dan pelaporan (misal : *spreadsheet* dan *pivot table*), analisis multidimensi (*dimensional summary*); OLAP (*Online Analytc Processing*) serta analisis statistik. Lalu bersifat Penemuan (*discovery driven*) bersifat transparan dan menjawab pertanyaan “mengapa?”.

Tahapan data mining dibagi menjadi bagian-bagian yaitu :

1. Pembersihan data (*data cleaning*)
2. Integrasi data (*data integration*)
3. Seleksi data (*data selection*)
4. Transformasi data (*data transformation*)
5. Proses *mining*
6. Evaluasi pola (*pattern evaluation*)
7. Pengetahuan (*knowledge presentation*)

## 2. *Clustering*

(Prasetyo Eko, 2014, p.186) mengatakan bahwa *Clustering* adalah teknik menemukan sekelompok data dari pemecahan atau pemisahan sekumpulan data menurut karakteristik tertentu yang telah ditentukan. Dalam pengelompokan tersebut nilai label nya belum diketahui sehingga diharapkan setelah melakukan pengelompokan data dapat diketahui label dari data tersebut. Metode *clustering* juga sering disebut tahapan awal sebelum melakukan metode lain seperti klasifikasi.

*Cluster analysis* adalah mengelompokan data objek pada informasi yang mirip atau memiliki kesamaan antara satu dengan yang lainnya, tujuannya agar dapat menemukan kelompok yang berkualitas seperti kelompok yang merupakan objek-objek yang mirip atau memiliki hubungan satu sama lain dan sebaliknya yaitu kelompok yang tidak berhubungan dengan objek dalam kelompok yang lain. *Clustering* cocok digunakan untuk menjelajahi data. Jika ada banyak kasus tapi tidak ada pengelompokan yang jelas, algoritma *clustering* dapat digunakan untuk mencari pengelompokan dari data tersebut. *Clustering* juga dapat berguna sebagai *data-preprocessing* yaitu langkah untuk mengidentifikasi kelompok-kelompok yang berhubungan dalam membangun model.

## 3. K-Means

(Prasetyo Eko, 2014, p.189), mengatakan bahwa Algoritma K-Means merupakan algoritma pengelompokan iteratif yang melakukan partisi set data ke dalam jumlah *K cluster* yang sudah ditetapkan di awal. Algoritma K-Means sederhana untuk di implementasikan dan dijalankan, relative cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, K-Means menjadi salah satu algoritma yang paling penting dalam bidang data *mining*. K-Means termasuk dalam *partitioning clustering* yaitu setiap data harus masuk dalam *cluster* tertentu dan memungkinkan bagi setiap data yang termasuk dalam *cluster* tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah, ke *cluster* yang lain. Algoritma K-Means sangat terkenal karena kemudahan dan kemampuannya untuk mengklasifikasikan data.

Langkah-langkah metode *k-means*:

1. Menentukan *k* sebagai jumlah *cluster* yang ingin dibentuk.
2. Menentukan nilai secara acak untuk pusat *cluster* awal *centroid* sebanyak *k*.
3. Menghitung jarak setiap data yang dimasukkan terhadap masing-masing *centroid* menggunakan rumus *Euclidian Distance* hingga ditemukan jarak

yang paling dekat antara setiap data dengan *centroid*. Berikut ini adalah persamaan *Euclidian Distance*:

$$(x_i, y_j) = \sqrt{(x_i - y_j)^2} \quad (2-1)$$

Keterangan:

$x_i$  = Data kriteria

$y_j$  = *Centroid* pada *cluster* j

4. Mengelompokkan setiap data berdasarkan kedekatannya dengan *centroid* (jarak terkecil).
5. Memperbaharui nilai *centroid*. Nilai *centroid* yang baru didapatkan dari rata-rata *cluster* yang bersangkutan dengan menggunakan rumus:

$$y_j(t+1) = \frac{1}{N_{sj}} \sum_{j \in N_{sj}} x_j \quad (2-2)$$

Keterangan:

$(t+1)$  = *Centroid* baru pada iterasi ke  $t + 1$

$N_{sj}$  = Banyaknya data pada *cluster* j

6. Melakukan perulangan dari langkah 2 hingga 5 sampai anggota tiap *cluster* tidak ada yang berubah.
7. Jika langkah 6 telah terpenuhi, maka nilai pusat *cluster* pada perulangan terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

**Contoh kasus:**

BPR ABC memiliki data nasabah yang pernah memperoleh kredit, data berupa jumlah rumah dan mobil yang dimiliki pelanggan yang disajikan pada Tabel 2.1.

**Tabel 2.1 Data Nasabah**

Nasabah	Jumlah Rumah	Jumlah Mobil
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

Clustering yang diharapkan mampu menghasilkan kelompok nasabah yang memenuhi sifat seperti, nasabah yang jumlah rumah dan mobilnya hampir sama akan berada pada

kelompok nasabah yang sama. Nasabah yang jumlah rumah dan mobilnya cukup berbeda akan berada pada kelompok yang berbeda.

Berikut langkah-langkah *Clustering* menggunakan algoritma *K-Means*:

Langkah 1: Tentukan jumlah cluster yang diinginkan (missal k=3)

Langkah 2: Pilih centroid awal secara acak. Pada langkah ini secara acak akan dipilih 3 buah data sebagai centroid, misalnya: data (B,E,F)

M1 = (3,3) ,M2 = (1,2) ,M3 = (4,2)

Langkah 3: Hitung jarak dengan centroid.Pada langkah ini setiap data akan ditentukan centroid terdekatnya, dan data tersebut akan ditetapkan sebagai anggota kelompok yang terdekat dengan centroid.

Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sebagai berikut: Data: (1,3), centroid M1: (3,3), centroid M2: (1,2), centroid M3: (4,2)

$$DM1 = \sqrt{(1 - 3)^2 + (3 - 3)^2} = 2$$

$$DM2 = \sqrt{(1 - 1)^2 + (3 - 2)^2} = 1$$

$$DM3 = \sqrt{(1 - 4)^2 + (3 - 2)^2} = 3.162$$

Hasil perhitungan terkait hitung jarak dengan centroid dapat dilihat pada Tabel 2.2 di mana perhitungan jarak titik dengan centroid dapat diketahui sehingga jarak terdekat dapat diketahui.

**Tabel 2.2 Hasil Perhitungan Jarak**

Nasabah	Jarak ke centroid cluster1	Jarak ke centroid cluster 2	Jarak ke centroid cluster 3	Jarak terdekat
A	2	1	3.162	C2
B	0	2.236	1.414	C1
C	1	3.162	1	C3
D	2	4.123	1.414	C3
E	2.236	0	3	C2
F	1.414	3	0	C3
G	2.828	1	3.162	C2
H	2.236	1.414	2.236	C2

Dari Tabel 2.2 didapatkan keanggotaan nasabah sebagai berikut.

Cluster 1 = {B}, Cluster 2 = {A,E,G,H}, Cluster 3 = {C,D,F}

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*):

Karena centroid  $M1=(3,3)$ ,  $M2=(1,2)$ ,  $M3=(4,2)$

$$d(m1,m2) = \sqrt{(3 - 1)^2 + (3 - 2)^2} = 2.236$$

$$d(m1,m3) = \sqrt{(3 - 4)^2 + (3 - 2)^2} = 1.414$$

$$d(m2,m3) = \sqrt{(1 - 4)^2 + (2 - 2)^2} = 3$$

$$BCV = d(m1,m2) + d(m1,m3) + d(m2,m3) = 2.236 + 1.414 + 3 = 6,650$$

Dalam hal ini  $d(m_i,m_j)$  menyatakan *Euclidean Distance* dari  $m$  ke  $m_j$

Sementara untuk menghitung WCV yaitu dengan memilih jarak terkecil antara data dengan centroid pada masing-masing cluster dapat dilihat pada

Tabel 2.3.

**Tabel 2.3 Jarak Antara Data Dengan Centroid**

Nasabah	Jarak ke centroid terkecil
A	1
B	0
C	1
D	1.414
E	0
F	0
G	1
H	1.414

$WCV=1^2+0^2+1^2+1.414^2+0^2+0^2+1^2+1.414^2 = 7$  Sehingga besar rasio  $BCV/WCV = 6.650/7 = 0.950$  Karena langkah ini merupakan iterasi 1 maka lanjutkan ke langkah berikutnya.

Langkah 4: Pembaruan centroid dengan menghitung rata-rata nilai pada masing-masing cluster. Tabel 2.4 menunjukkan terkait hasil pembagian *Cluster*.

**Tabel 2.4 Pembagian Cluster**

<b>Cluster 1</b>		
Nasabah	Jumlah Rumah	Jumlah Mobil
B	3	3
Mean	3	3
<b>Cluster 2</b>		
Nasabah	Jumlah Rumah	Jumlah Mobil
A	1	3
E	1	2

G	1	1
H	2	1
Mean	1.25	1.75
<b>Cluster 3</b>		
Nasabah	Jumlah Rumah	Jumlah Mobil
C	4	3
D	5	3
F	4	2
Mean	4.33	2.67

Sehingga didapatkan centroid baru yaitu:  $m_1=(3,3)$ ,  $m_2=(1.25,1.75)$ ,  $m_3=(4.33,2.67)$

Langkah 3: (iterasi-2) kembali ke langkah 3, jika masih ada data yang berpindah cluster atau jika nilai centroid di atas nilai ambang, atau jika nilai pada fungsi obyektif yang digunakan masih di atas ambang. Selanjutnya pada langkah ini dilakukan penempatan lagi data dalam centroid terdekat sama seperti yang dilakukan di langkah-3, untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sebagai berikut: Data : (1,3),  $m_1=(3,3)$ ,  $m_2=(1.25,1.75)$ ,  $m_3=(4.33,2.67)$

$$dm_1 = \sqrt{(1-3)^2 + (3-3)^2} = 2 \quad dm_2 = \sqrt{(1-1.25)^2 + (3-1.75)^2} = 1.275 \quad dm_3 = \sqrt{(1-4.33)^2 + (3-2.67)^2} = 3.350$$

hasil perhitungan terkait jarak centroid dengan titik nasabah dapat dilihat pada Tabel 2.5 terkait jarak terdekat.

**Tabel 2.5 Hasil Perhitungan Jarak Centroid**

Nasabah	Jarak ke centroid cluster1	Jarak ke centroid cluster2	Jarak ke Centroid cluster3	Jarak terdekat
A	2	1.275	3.350	C2
B	0	1.768	1.374	C1
C	1	3.021	0.471	C3
D	2	3.953	0.745	C3
E	2.236	0.354	3.399	C2
F	1.414	2.813	0.745	C3
G	2.828	0.791	3.727	C2
H	2.236	1.061	2.867	C2

Dari tabel 2.6 didapatkan keanggotaan nasabah sebagai berikut:

Cluster 1= {B}, Cluster 2= {A,E,G,H}, Cluster 3= {C,D,F}

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*):

$$BCV = d(m1,m2) + d(m1,m3) + d(m2,m3) = 6,741$$

WCV =  $1.275^2 + 0^2 + 0.471^2 + 0.745^2 + 0.354^2 + 0.745^2 + 0.791^2 + 1.064^2 = 4.833$  Sehingga besar rasio  $BCV/WCV = 6.741/4.833 = 1.394$  Bila dibandingkan maka rasio sekarang (1.394) lebih besar dari rasio sebelumnya (0.950) oleh karena itu algoritma dilanjutkan ke langkah berikutnya.

Langkah ke-4 – iterasi 3 : Pada langkah ini dilakukan pembaruan centroid lagi. Hasil terkait pembagian *cluster* dapat dilihat pada Tabel 2.6 dari masing-masing *Cluster*.

**Tabel 2.6 Pembagian Cluster**

<b>Cluster 1</b>		
Nasabah	Jumlah Rumah	Jumlah Mobil
B	3	3
Mean	3	3
<b>Cluster 2</b>		
Nasabah	Jumlah Rumah	Jumlah Mobil
A	1	3
E	1	2
G	1	1
H	2	1
Mean	1.25	1.75
<b>Cluster 3</b>		
Nasabah	Jumlah Rumah	Jumlah Mobil
C	4	3
D	5	3
F	4	2
Mean	4.33	2.67

Langkah ketiga iterasi 3

Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sebagai berikut: Data nasabah A: (1,3),  $m1=(3,3)$ ,  $m2=(1.25,1.75)$ ,  $m3=(4.33,2.67)$

$$dm1 = \sqrt{(1-3)^2 + (3-3)^2} = 2 \quad dm2 =$$

$$\sqrt{(1-1.25)^2 + (3-1.75)^2} = 1.275 \quad dm3 =$$

$$\sqrt{(1-4.33)^2 + (3-2.67)^2} = 3.350$$

Tabel 2.7 menunjukkan hasil dari perhitungan jarak titik tengah dari seluruh nasabah.

**Tabel 2.7 Hasil Perhitungan Jarak Centroid**

Nasabah	Jarak ke centroid cluster1	Jarak ke centroid cluster2	Jarak ke centroid cluster3	Jarak terdekat
A	2	1.275	3.350	C2
B	0	1.768	1.374	C1
C	1	3.021	0.471	C3
D	2	3.953	0.745	C3
E	2.236	0.354	3.399	C2
F	1.414	2.813	0.745	C3
G	2.828	0.791	3.727	C2
H	2.236	1.061	2.867	C2

Dari tabel 2.7 didapatkan keanggotaan nasabah sebagai berikut:

Cluster 1 = {B}, Cluster 2 = {A,E,G,H}, Cluster 3 = {C,D,F}

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*):

$$BCV = d(m1,m2) + d(m1,m3) + d(m2,m3) = 6,741$$

$$WCV = 1.275^2 + 0^2 + 0.471^2 + 0.745^2 + 0.354^2 + 0.745^2 + 0.791^2 + 1.064^2 = 4.833$$

$$\text{Sehingga besar rasio } BCV/WCV = 6.741/4.833 = 1.394$$

Bila dibandingkan maka rasio sekarang (1.394) sudah tidak lagi lebih besar dari rasio sebelumnya (1.394) oleh karena itu algoritma akan dihentikan. Metode K-means perhitungannya akan terhenti bila tidak ada pergerakan dari anggota clusternya, atau pengelompokan dari iterasi yang sebelumnya dan yang sekarang sedang dikelompokkan hasilnya sama

#### 4. Pengembangan Sistem SDLC

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 (Systems development life cycle-SDLC) adalah aplikasi dari pendekatan sistem bagi pengembangan suatu sistem informasi.

SDLC tradisional tidak membutuhkan waktu lama bagi seorang pengembang sistem yang pertama untuk mengetahui bahwa terdapat beberapa tahapan pekerjaan pengembangan yang perlu dilakukan dalam urutan tertentu jika suatu proyek ingin memiliki kemungkinan berhasil yang paling besar. Tahapan-tahapan tersebut yaitu perencanaan, analisis desain implementasi

Penggunaan Proyek direncanakan dan sumber-sumber daya yang dibutuhkan untuk melakukan pekerjaan 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 mengilustrasikan sifat melingkar dari siklus hidup. Ketika sebuah sistem telah melampaui masa manfaatnya dan harus diganti, satu siklus hidup baru akan dimulai dengan diawali oleh tahap perencanaan.

Mudah untuk melihat bagaimana SDLC tradisional dapat dikatakan sebagai suatu aplikasi dari pendekatan sistem. Masalah akan didefinisikan dalam tahap-tahap 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 (Raymond McLeod and George, 2007, p.199 - 200).



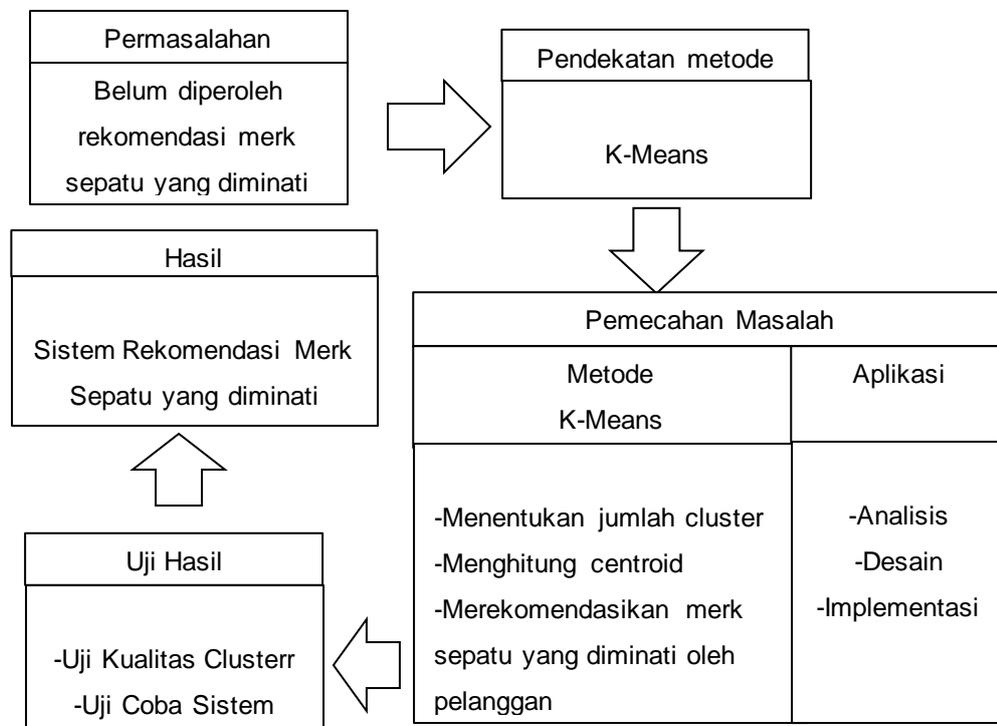
**Gambar 2.1 Pola melingkar dari siklus hidup sistem  
(Sumber: Raymond Mc Leod and George, 2007, p. 200)**

Tahap-tahap SDLC tradisional ada kekurangannya. Ketika sistem tumbuh dalam ukuran dan kompleksitas, menjadi tidak mungkin untuk membuat satu tembakan melewati tahapan. Pengembang selalu mengulangi dan mengulangi hal-hal yang muncul dengan sistem yang memuaskan pengguna. Juga, proyek cenderung berlarut-larut selama berbulan-bulan dan bertahun-tahun, dan hampir selalu melebihi anggaran mereka. Menanggapi keterbatasan ini, pengembang sistem memutuskan untuk menerapkan teknik yang telah terbukti efektif dalam pengejaran lainnya, seperti desain mobil - penggunaan prototipe. Seperti yang

diterapkan pada pengembangan sistem, prototipe adalah versi dari sistem potensial yang memberikan para pengembang dan pengguna masa depan dengan gagasan tentang bagaimana sistem dalam bentuk yang lengkap akan berfungsi. Proses menghasilkan prototipe disebut prototyping. Idenya adalah untuk menghasilkan prototipe secepat mungkin, mungkin dalam semalam, dan mendapatkan umpan balik pengguna yang akan memungkinkan prototipe ditingkatkan lagi dengan sangat cepat (Raymond Mc Leod and George, 2007).

### C. Kerangka Pemikiran

Berdasarkan dukungan landasan teoritis yang diperoleh dari eksplorasi teori yang dijadikan rujukan penelitian, maka dapat disusun kerangka pemikiran pada gambar 2.2



**Gambar 2.2 Kerangka Pemikiran**

Kerangka pemikiran pada gambar 2.2 dapat dijelaskan sebagai berikut :

1. Komponen permasalahan mencakup fenomena yaitu sulit dalam menentukan merk sepatu yang diminati pelanggan toko.
2. Pendekatan metode yang digunakan yaitu dengan K-means.
3. Dalam pemecahan masalah dibagi menjadi dua, yaitu pemecahan dalam sisi metode K-means dan juga dalam aplikasi.
4. Hasilnya yang di dapatkan adalah sebuah sistem rekomendasi merk sepatu yang diminati oleh pelanggan.

5. Pengujian hasil kualitas *cluster* dan uji sistem.

#### **D. Hipotesis**

Hipotesis dalam penelitian ini yaitu metode K-Means diduga dapat memberikan rekomendasi merk sepatu yang diminati pelanggan.