BAB 2

LANDASAN TEORI

2.1 Tinjauan Studi

Dalam penelitian yang dilakukan sebelumnya dengan judul Penerapan Data Mining Untuk Menentukan Strategi Promosi Universitas PGRI Yogyakarta Menggunakan Algoritma K-means Clustering. Penelitian ini melakukan strategi untuk promosi Universitas PGRI Yogyakarta dengan menggunakan K-means clustering agar mengetahui persebaran wilayah berdasarkan indeks prestasi mahasiswa. Dengan menerapkan metode algoritma K-Means Clustering menggunakan dua aplikasi yaitu Microsoft Excel 2019 dan Sistem Aplikasi K-Means Clustering [7].

Dalam Penelitian selanjutnya dengan judul Implementasi Data Mining Menggunakan Algoritma K-Means Clustering Untuk Mengetahui Pola Pemilihan 8 Program Studi IAIN Salatiga. Penelitian ini bertujuan untuk mencari pola pemilihan program studi yang akan dilakukan pada mahasiswa IAIN Salatiga. Dengan pengolahan data mining menggunakan algoritma k-means clustering untuk mengetahui pola pemilihan progam studi IAIN Salatiga, dilakukan dengan menggunakan dua aplikasi, yang pertama dengan menggunakan aplikasi pengolah angka Microsoft Excel dan dengan menggunakan apalikasi yang peneliti rancang dengan menggunakan bahasa pemrograman PHP dan Database MySQL. Hasil dari proses k-means clustering dengan menggunakan data mahasiswa IAIN Salatiga S1 angkatan 2016 terbagi ke dalam lima cluster, dimana cluster pertama berisikan tentang program studi yang paling diminati oleh mahasiswa baru IAIN Salatiga, cluster kedua berisikan prodi dengan peminat lebih sedikit dengan dari cluster pertama dan lebih banyak di bandingkan dengan cluster tiga, empat dan lima, pada cluster tiga berisikan program studi yang diminati, sedangkan pada cluster empat dan lima berisikan program studi yang kurang diminati oleh mahasiswa baru [8].

2.2 Tinjauan Pustaka

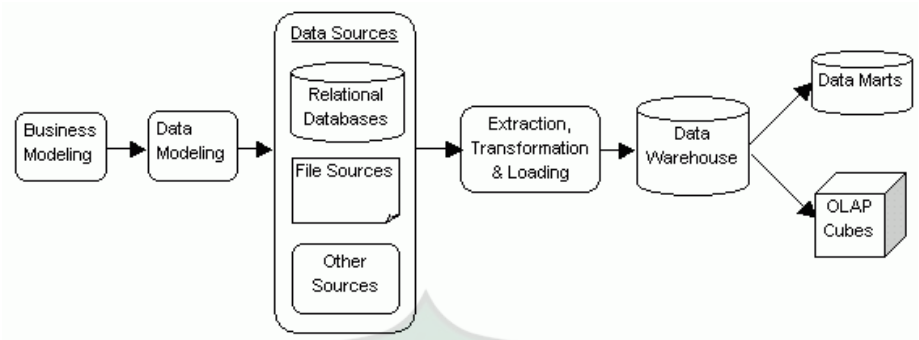
Dalam landasan teori ini akan dijelaskan secara garis besar teori-teori yang menjadi dasar atau acuan dalam penelitian ini:

2.2.1 Bussines Intelligence

Business Intelligence (BI) adalah sebuah proses ekstraksi data operasional organisasi atau perusahaan, kemudian mengumpulkannya kedalam sebuah data warehouse. Sebuah data warehouse dirancang untuk mendukung sebuah proses lanjutan dalam rangka mendapatkan informasi berharga menggunakan teknik data mining. Konsep BI menekankan pada penerapan 5 pendayagunaan informasi untuk keperluan spesifik bisnis, masing-masing adalah sebagai berikut [1]:

1) Data sourcing.
2) Data analysis.
3) Situation awareness.
4) Risk analysis.
5) Decission support.

Dalam membuat business modelling digunakan business model dan diagram yang memberikan informasi secara grafis bagi anggota suatu organisasi atau perusahaan memahami dan mengkomunikasikan *business rule* dan proses-proses bisnisnya [1].



Gambar 2.1 Business Intelligence Environment [9]

Seorang kepala biro marketing, dia dapat melakukan kampanye pemasaran dengan segmentasi target yang jelas dan dapat diperhitungkan dalam menyumbang penerimaan mahasiswa baru bagi institusinya dengan mencari strategi yang tepat dalam melakukan promosi, dan masih banyak lagi evaluasi kinerja manajemen dari setiap divisi yang dilakukan pada suatu organisasi atau perusahaan.

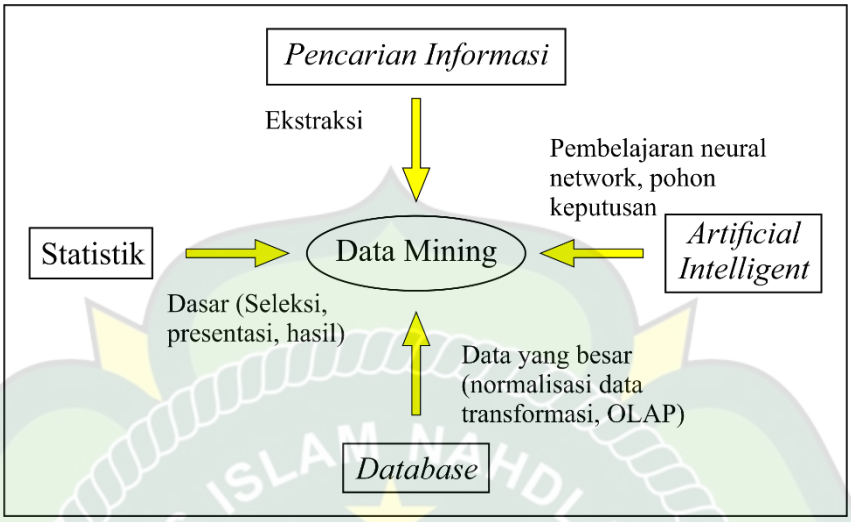
2.2.2 Pengertian Data Mining

Data mining adalah proses menganalisis data dan menemukan pola tersembunyi secara otomatis atau semi otomatis [10]. Pola atau hubungan digunakan sebagai alat untuk menjelaskan data dan memprediksi data yang baru [11]. Ada beberapa definisi lain dari *data mining*. *Data mining* adalah proses menganalisis data yang sangat besar untuk menemukan hubungan dan merepresentasikan data yang berguna dan dapat dipahami untuk pemiliknya [12].

Secara sederhana, *data mining* berarti menggali pengetahuan dari data yang berjumlah banyak [13].

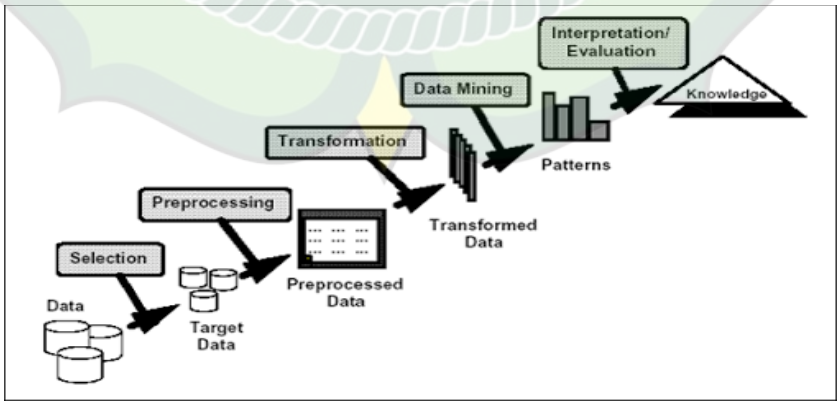
Salah satu kesulitan mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah. Gambar 2.2 merupakan bidang ilmu yang menjadi akar panjang dari data mining. Beberapa bidang ilmu tersebut seperti kecerdasan

buatan, *machine learning*, statistik, database, dan juga information retrieval [14].



Gambar 2.2 Bidang Ilmu Data Mining [15]

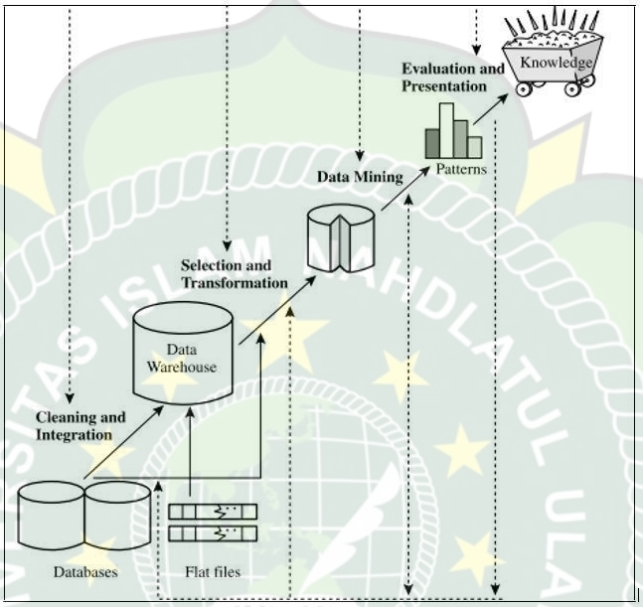
Data Mining terkadang dianggap sama dengan istilah KDD (Knowledge Discovery in Database). Namun, Data Mining adalah salah satu bagian dari KDD. Dibandingkan dengan KDD, Data Mining lebih dikenal di kalangan pelaku bisnis. Sebagai komponen dalam KDD, Data Mining berkaitan dengan ekstraksi dan perhitungan pola-pola yang telah dianalisis. Tahapan dalam proses KDD dijelaskan di bawah ini: [14]



Gambar 2.3 Tahapan Proses Knowledge Discovery in Database [14]

2.2.3 Tahapan Data Mining

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan knowledge base. Tahapan-tahapan tersebut, diantaranya ditunjukkan pada Gambar 2.3:



Gambar 2.4 Bidang Ilmu Data Mining [15]

Serangkaian proses tahapan data mining tersebut memiliki tahap sebagai berikut : [13]

1) Data Cleaning

   Data cleaning merupakan proses menghilangkan *noise*, data yang tidak konsisten, dan data yang tidak relevan.

2) Data Integration

   Data integration merupakan proses menggabungkan data dari berbagai data sumber (data source) ke dalam *database* yang akan digunakan untuk proses penggalian data.

3) Data Selection

Data selection merupakan proses pemilihan data yang digunakan untuk proses penggalian data.

4) Data Transformation

Data Transformation merupakan proses mentransformasikan dan mengkonsolidasikan data untuk digunakan dalam proses *mining*.

5) Data Mining

Data mining merupakan proses utama mencari pengetahuan atau pola dari informasi tersembunyi dari *database*.

6) Pattern Evaluation

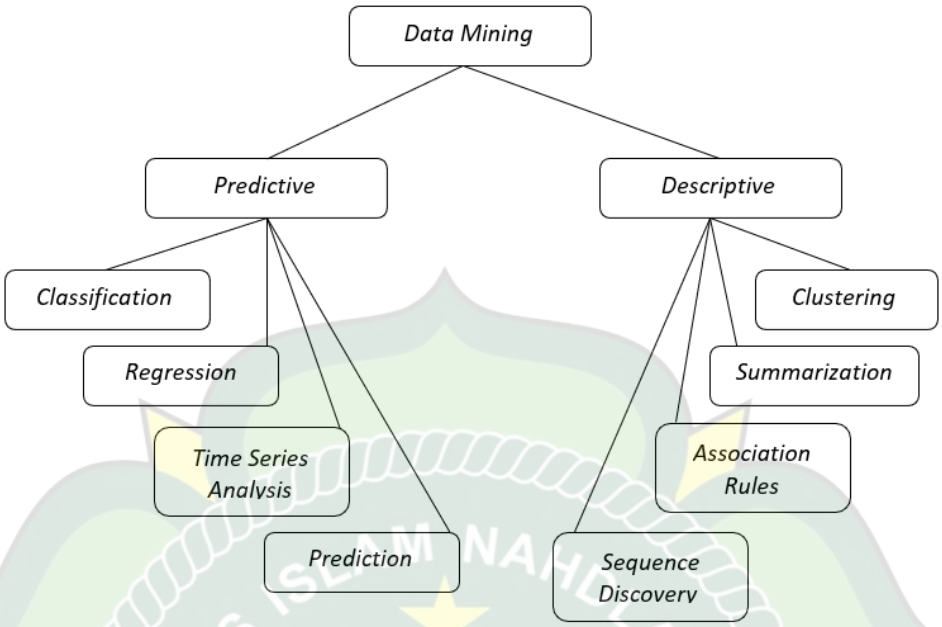
Pattern Evaluation merupakan proses mengidentifikasi pola yang telah didapat.

7) Knowledge Presentation

Knowledge Presentation merupakan visualisasi dan presentasi pengetahuan atau pola yang telah didapat kepada pengguna.

2.2.4 Teknik-Teknik Data Mining

Menurut Ahmed, teknik data mining biasanya terbagi dalam dua kategori, prediksi dan deskripsi. Teknik prediksi menggunakan data historis untuk menyimpulkan sesuatu tentang kejadian di masa depan. Sedangkan teknik deskripsi bertujuan untuk menemukan pola dalam data yang menyediakan beberapa informasi tentang hubungan interval yang tersembunyi.



Gambar 2.5 Teknik Data Mining [16]

Menurut Kumar dan Saurabh, terdapat beberapa teknik yang digunakan dalam *data mining*, yaitu: [16]

1. Classification

   Klasifikasi adalah teknik yang paling umum diterapkan pada *data mining*. Pendekatan ini sering menggunakan keputusan pohon (*decision tree*) atau *neural network* berbasis algoritma klasifikasi.

2. Clustering

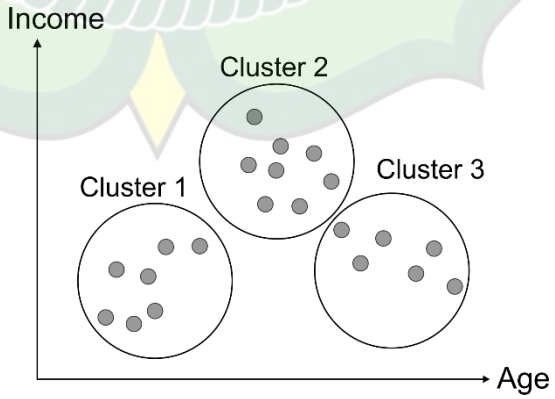
   Clustering bisa dikatakan sebagai identifikasi kelas objek yang memiliki kemiripan. Dengan menggunakan teknik *clustering* kita bisa lebih lanjut mengidentifikasi kepadatan dan jarak daerah dalam objek ruang dan dapat menemukan secara keseluruhan pola distribusi dan korelasi antara atribut.

Pendekatan klasifikasi secara efektif juga dapat digunakan untuk membedakan kelompok atau kelas objek.

Metode ini digunakan untuk mengidentifikasi kelompok alami dari sebuah kasus yang didasarkan pada sebuah kelompok atribut, mengelompokkan data yang memiliki kemiripan atribut. Gambar 2.5 Clustering ini menunjukkan kelompok data pelanggan sederhana yang berisi dua atribut, yaitu *Age* (Umur) dan *Income* (Pendapatan). Algoritma *Clustering* mengelompokkan kelompok data kedalam tiga segment berdasarkan kedua atribut ini [17].

a) Cluster 1 berisi populasi berusia muda dengan pendapatan rendah.
b) Cluster 2 berisi populasi berusia menengah dengan pendapatan yang lebih tinggi.
c) Cluster 3 berisi populasi berusia tua dengan pendapatan yang relatif rendah.

Clustering adalah metode Data Mining yang *Unsupervised*, karena tidak ada satu atribut yang digunakan untuk memandu proses pembelajaran, jadi seluruh atribut input diperlakukan sama.



Gambar 2.6 Clustering [17]

3. Regression

   Metode Regression mirip dengan metode Classification, Metoda regression bertujuan untuk mecari pola dan menentukan sebuah nilai numerik. Regression digunakan untuk memecahkan suatu permasalahan, contohnya untuk memperkirakan kecepatan angin berdasarkan temperatur, tekanan udara, dan kelembaban.

4. Decision trees

   Decision trees atau pohon keputusan merupakan model prediktif yang dapat digambarkan seperti pohon, dimana setiap node didalam struktur pohon tersebut mewakili sebuah pertanyaan yang digunakan untuk menggolongkan data. Struktur ini dapat digunakan untuk membantu memperkirakan kemungkinan nilai setiap atribut data.

2.2.5 Metode Data Mining

Salah satu metode data mining adalah model Cross-Standard Industry for Data Mining (CRISP-DM) yang terdiri dari 6 fase, yaitu [1]:

1) Fase pemahaman bisnis (Business Un derstanding)

   Pada tahap ini berfokus pada pemahaman mengenai tujuan dari proyek dan kebutuhan secara persepktif bisnis, kemudian mengubah hal tersebut menjadi sebuah permasalahan *data mining* dan rencana awal untuk mencapai tujuan tersebut. Kegiatan yang dilakukan antara lain: menentukan tujuan dan persyaratan dengan jelas secara keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah *data mining*, dan selanjutnya mempersiapkan strategi awal untuk mencapai tujuan tersebut.

2) Fase pemahaman data (Data Understanding)

Pada tahap ini dilakukan pengumpulan terhadap data, lalu kemudian mempelajari data tersebut dengan tujuan untuk mengenal data, melakukan identifikasi dan mengetahui kualitas dari data, serta mendeteksi subset yang menarik dari data yang dapat dijadikan hipotesa bagi informasi yang tersembunyi.

3) Fase pengolahan data (Data Preparation)

Pada tahap ini dilakukan persiapan mengenai data yang akan digunakan pada tahap berikutnya. Kegiatan yang dilakukan antara lain: memilih kasus dan parameter yang akan dianalisis (*Select Data*), melakukan transformasi terhadap parameter tertentu (*Transformation*), dan melakukan pembersihan data agar data siap untuk tahap *modelling* (*Cleaning*). Data *preprocessing* bertujuan untuk mendapatkan data yang bersih dan siap untuk digunakan dalam penelitian. Tahapan yang dikerjakan adalah melakukan pengabaian atribut pada data mentah yang dianggap tidak relevan dengan hasil pengujian dan perubahan terhadap nilai data bahkan tipe data pada atribut *dataset* dengan tujuan untuk mempermudah pemahaman terhadap isi *record* dengan memperhatikan konsistensi data, *missing value*, dan *redundancy* pada data.

4) Fase pemodelan (Modelling)

Pada tahap ini dilakukan penentuan terhadap teknik *data mining*, alat bantu *data mining*, dan algoritma data mining yang akan diterapkan. Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma *data mining* tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik data mining tertentu, dapat kembali ke tahap persiapan data.

5) Fase Evaluasi (Evaluation)

Pada tahap ini dilakukan pengujian terhadap model-model yang dikomparasi untuk mendapatkan informasi model yang paling akurat. Evaluasi dan validasi menggunakan metode *confusion matrix* dan kurva ROC.

6) Fase Penyebaran (Deployment)

Setelah pembentukan model dan dilakukan analisa dan pengukuran pada tahap sebelumnya, selanjutnya pada tahap ini diterapkan model yang paling akurat dengan memakai data baru diluar data *training* dan data *testing*.

2.2.6 Clustering

Menurut Deka, *Clustering* merupakan salah satu teknik *data mining* yang digunakan untuk mendapatkan kelompok-kelompok dari objek-objek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data atau objek ke dalam *cluster* atau grup sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin. *Clustering* melakukan pengelompokkan data yang didasarkan pada kesamaan antar objek, oleh karena itu klasterisasi digolongkan sebagai metode *unsupervised learning*. Menurut Oyelade, *clustering* dapat dibagi menjadi dua, yaitu *hierarchical clustering* dan *non-hierarchical clustering*.

Hierarchical clustering adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster*.

Berbeda dengan metode *hierarchical clustering*, metode *non-hierarchical clustering* justru dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (dua *cluster*, tiga *cluster*, atau lain sebagainya). Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan *K-Means Clustering* [16].

2.2.7 Algoritma K-Means Clustering

K-Means merupakan salah satu algoritma *clustering*. Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan *supervised learning* yang menerima masukan berupa vektor (x_1, y_1), (x_2, y_2), …, (x_i, y_i), di mana x_i merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk x_i. Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk *unsupervised learning*. Masukan yang diterima adalah data atau objek dan *k* buah kelompok (*cluster*) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam *k* buah kelompok tersebut. Pada setiap *cluster* terdapat titik pusat (*centroid*) yang merepresentasikan *cluster* tersebut.

K-Means ditemukan oleh beberapa orang yaitu Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967). Ide dari *clustering* pertama kali ditemukan oleh Lloyd pada tahun 1957, namun hal tersebut baru dipublikasi pada tahun 1982.

K-Means merupakan metode *klasterisasi* yang paling terkenal dan banyak digunakan di berbagai bidang karena sederhana, mudah diimplementasikan, memiliki kemampuan untuk mengklaster data yang besar, mampu menangani *data outlier*, dan kompleksitas waktunya linear O(nKT) dengan n adalah jumlah dokumen, K adalah jumlah kluster, dan T

adalah jumlah iterasi. K-Means merupakan metode pengklasteran secara partitioning yang memisahkan data ke dalam kelompok yang berbeda.

Pada pengukuran jarak antara setiap objek data dan cluster centroid dapat menggunakan perhitungan seperti *euclidean distance*, *manhattan distance*, dan *minkowsky distance*. Adapun karakteristik dari algoritma K-Means salah satunya adalah sangat sensitif dalam penentuan titik pusat awal klaster karena K-Means membangkitkan titik pusat klaster awal secara random. Pada saat pembangkitan awal titik pusat yang random tersebut mendekati solusi akhir pusat klaster, K-Means mempunyai posibilitas yang tinggi untuk menemukan titik pusat klaster yang tepat. Sebaliknya, jika awal titik pusat tersebut jauh dari solusi akhir pusat klaster, maka besar kemungkinan ini menyebabkan hasil pengklasteran yang tidak tepat. Akibatnya K-Means tidak menjamin hasil pengklasteran yang unik. Inilah yang menyebabkan metode K-Means sulit untuk mencapai optimum global, akan tetapi hanya minimum lokal. Selain itu, algoritma K-Means hanya bisa digunakan untuk data yang atributnya bernilai numerik.

Metode K-Means Clustering adalah proses untuk mengelompokkan data ke dalam sebuah cluster dengan titik pusat yang berbeda-beda setiap cluster. Proses K-Means Clustering tersebut meliputi 5 proses, yaitu [18]:

1) Menentukan Titik Pusat Cluster

   Menentukan titik pusat cluster adalah langkah awal untuk proses K-Menas Clustering. Fungsi proses ini adalah untuk menentukan titik awal sebagai patokan untuk mencari jarak antara data ke cluster yang sudah ditentukan. Titik awal pusat cluster disebut juga dengan centroid. Untuk menentukan titik pusat setiap cluster bisa dilakukan dengan mencari rata-rata dari data yang akan diolah ataupun sesuai keinginan.

2) Menghitung Jarak Data ke Setiap Cluster

Setelah menentukan titik pusat di setiap cluster proses selanjutnya adalah menghitung jarak antara data ke setiap cluster yang sudah dibentuk. Rumus untuk mencari jarak (distance) dari satu cluster adalah

Rumus 2.1 Rumus Mencari Jarak Data ke Setiap Cluster

$$\sqrt{(X_i - X_{avg})^2 + (Y_i - X_{avg})^2 + (Z_{ki} - Z_{kj})^2}$$

Dimana:

X_i : Data pertama (diambil dari atribut pertama)

X_{avg} : Titik pusat cluster / centroid untuk atribut pertama

Y_i : Data kedua (diambil dari atribut kedua)

Y_{avg} : Titik pusat cluster / centroid untuk atribut kedua

Z_i : Data ketiga (diambil dari atribut ketiga)

Z_{avg} : Titik pusat cluster / centroid untuk atribut pertama

Rumus tersebut adalah rumus untuk menentukan jarak dari satu baris data ke satu cluster tertentu.

3) Mengalokasikan Data kedalam Cluster

Setelah mendapatkan jarak antara setiap data ke setiap cluster yang terbentuk maka proses selanjutnya adalah clustering atau mengelompokkan dan mengalokasikan data ke dalam cluster. Untuk mengelompokkan data ke dalam cluster ini cukup dilihat dari jarak terdekat dari setiap cluster. Apabila jarak yang didapatkan dari suatu data adalah dengan urutan paling kecil di setiap cluster, maka data termasuk kedalam cluster pertama.

4) Menentukan Titik Pusat Cluster Baru

Setelah mengalokasikan data ke dalam cluster yang terbentuk maka proses selanjutnya adalah menentukan titik pusat cluster baru dengan cara yang sama yaitu mencari rata-rata di setiap atribut data. Tetapi dalam perhitungan kali ini sedikit berbeda dari yang pertama, apabila proses pertama adalah mencari rata-rata dari semua atribut data maka untuk menentukan titik pusat cluster baru ini data yang digunakan sesuai dengan clusternya masing-masing. Apabila data yang masuk ke cluster satu adalah data pertama, ketiga, kelima, dan keenam, maka rata-rata yang dicari hanya menggunakan data pertama, ketiga, kelima, dan keenam.

5) Memverifikasi Titik Pusat Cluster

Setelah mendapatkan titik pusat cluster baru maka proses selanjutnya adalah memverifikasi titik pusat cluster baru tersebut dengan titik pusat cluster yang lama. Apabila hasil titik pusat cluster baru yang didapat sama dengan titik pusat cluster yang lama, maka proses K-Means sudah selesai dan hasil dari proses K-Means Clustering sudah didapatkan dan data yang diklasifikasi sudah tidak bisa berubah-ubah lagi. Tetapi jika hasil titik pusat cluster baru yang didapat berbeda dari titik pusat cluster yang lama, maka proses K-Means tetap dilanjutkan dan mulai lagi dari proses kedua atau menghitung jarak data ke setiap cluster.

Menurut Daniel dan Eko, Langkah-langkah algoritma *K-Means* adalah sebagai berikut [19]:

a) Pilih secara acak k buah data sebagai pusat cluster.

b) Jarak antara data dan pusat cluster dihitung menggunakan Euclidian Distance. Untuk menghitung jarak semua data ke

setiap titik pusat cluster dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}$$

dimana:

D *(i,j)* = Jarak data ke *i* ke pusat *cluster j*

X_{ki} = Data ke *i* pada atribut data ke *k*

X_{kj} = Titik pusat ke *j* pada atribut ke *k*

c) Data ditempatkan dalam cluster yang terdekat, dihitung dari tengah cluster.

d) Pusat cluster baru akan ditentukan bila semua data telah ditetapkan dalam cluster terdekat.

e) Proses penentuan pusat cluster dan penempatan data dalam cluster diulangi sampai nilai centroid tidak berubah lagi.

Berikut ini adalah contoh penerapan algoritma K-Means [16]:

Table 2.1 Data Mahasiswa

No	Nama	Jurusan	Kota Asal	IPK
1	Ade Supryan Stefanus	IS	Jakarta	3,16
2	Adelina Ganardi Putri Hardi	ACC	Semarang	3,22
3	Adeline Dewita	BF	Bekasi	3,29
4	Adiputra	IB	Jakarta	2,83
5	Afrieska Laura Trisyana	PR	Jakarta	3,15
6	Agam Khalilullah	IB	Banda Aceh	3,25
7	Agus Mulyana Jungjungan	IB	Bogor	3,43
8	Agusman	PR	Bekasi	3,06

9	Aidil Friadi	BF	Banda Aceh	3,36
10	Ajeng Putri Ariandhani	ACC	Bandung	3,28

2.2.8 Transformasi Data

Agar data di atas dapat diolah dengan menggunakan metode *k-means clustering*, maka data yang berjenis data *nominal* seperti kota asal dan jurusan harus diinisialisasikan terlebih dahulu dalam bentuk angka.

Table 2.2 Inisialisasi Data Wilayah Kota Asal

Wilayah	Frekuensi	Inisial
Jakarta	84	1
Jawa Barat	82	2
Sumatera Utara	28	3
Sulawesi	14	4
Jawa Timur	13	5
Sumatera Selatan	13	6
Bali	8	7
Kalimantan	1	8

Table 2.3 Inisialisasi Data Jurusan

Jurusan	Singkatan	Frekuensi	Inisial
Accounting	ACC	46	1
Management, concentration in International Business	IB	37	2
Public Relation	PR	35	3
Management, concentration in	BF	28	4

Banking & Finance			
Industrial Engineering	IE	23	5
Information Technology	IT	20	6
Management, concentration in Marketing	MKT	18	7
Visual Communication Design	VCD	12	8
Management, concentration in Hotel & Tourism Management	HTM	9	9
Electrical Engineering	EE	6	10
Business Administration	BA	4	11
International Relations	IR	2	12
Management, concentration in Human Resources Management	HRM	1	13
Information System	IS	1	14
Management	MGT	1	15

2.2.9 Pengolahan data

Setelah semua data mahasiswa ditransformasi ke dalam bentuk angka, maka data-data tersebut telah dapat dikelompokan dengan menggunakan algoritma K-Means Clustering. Untuk dapat melakukan pengelompokan data-data tersebut menjadi beberapa cluster perlu dilakukan beberapa langkah, yaitu:

1) Tentukan jumlah cluster yang diinginkan. Dalam penelitian ini data-data yang ada akan dikelompokkan mejadi tiga cluster.
2) Tentukan titik pusat awal dari setiap cluster. Dalam penelitian ini titik pusat awal ditentukan secara random dan didapat titik pusat dari setiap cluster dapat dilihat pada tabel 2.4.

Table 2.4 Titik Pusat Awal Setiap *Cluster*

Titik Pusat awal	Nama	Jurusan	Kota Asal	IPK
Cluster 1	Dally Teguh Sesario	9	3	2,94
Cluster 2	Hervina Juliana	1	1	3,18
Cluster 3	Pascal Muhammadi	1	2	3,15

3) Tempatkan setiap data pada cluster. Dalam penelitian ini digunakan metode hard k-means untuk mengalokasikan setiap data ke dalam suatu cluster, sehingga data akan dimasukan dalam suatu cluster yang memiliki jarak paling dekat dengan titik pusat dari setiap cluster. Untuk mengetahui cluster mana yang paling dekat dengan data, maka perlu dihitung jarak setiap data dengan titik pusat setiap cluster. Sebagai contoh, akan dihitung jarak dari data mahasiswa pertama ke pusat cluster pertama:

$$D\,(1,1) = \sqrt{(14-9)^2 + (1-3)^2 + (3{,}16-2{,}94)^2} = 5{,}390$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster pertama adalah 5,390. Jarak data mahasiswa pertama ke pusat cluster kedua:

$$D\,(1,2) = \sqrt{(14-1)^2 + (1-1)^2 + (3{,}16-3{,}18)^2} = 13{,}000$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster kedua adalah 13. Jarak data mahasiswa pertama ke pusat cluster ketiga:

$$D\,(1,3) = \sqrt{(14-1)^2 + (1-2)^2 + (3{,}16-3{,}15)^2} = 13{,}038$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat cluster ketiga adalah 13.038.

Berdasarkan hasil ketiga perhitungan di atas dapat disimpulkan bahwa jarak data mahasiswa pertama yang paling dekat adalah dengan *cluster* 1, sehingga data mahasiswa pertama dimasukkan ke dalam *cluster* 1. Hasil perhitungan selengkapnya untuk 5 data mahasiswa pertama dapat di lihat pada tabel 2.5.

Table 2.5 Contoh Hasil Perhitungan Setiap Data ke Setiap *Cluster*

No	Nama	Jurusan	Kota Asal	IPK	Jarak Ke			Jarak terdekat ke *Cluster*
					C1	C2	C3	
1	Ade Supryan Stefanus	14	1	3,16	5,390	13,000	13,038	1
2	Adelina Ganardi Putri Hardi	1	5	3,22	8,251	4,000	3,001	3
3	Adeline Dewita	4	2	3,29	5,111	3,164	3,003	3
4	Adiputra	2	1	2,83	7,281	1,059	1,450	2
5	Afrieska Laura Trisyana	3	1	3,15	6,328	2,000	2,236	2

4) Setelah semua data ditempat ke dalam *cluster* yang terdekat, kemudian hitung kembali pusat *cluster* yang baru berdasarkan rata-rata angggota yang ada pada *cluster* tersebut.
5) Setelah didapatkan titik pusat yang baru dari setiap cluster, lakukan kembali dari langkah ketiga hingga titik pusat dari setiap cluster tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu cluster ke cluster yang lain.

Dalam contoh ini, iterasi clustering data mahasiswa terjadi sebanyak 7 kali iterasi. Pada iterasi ke-7 ini, titik pusat dari setiap cluster sudah tidak berubah dan tidak ada lagi data yang berpindah dari satu cluster ke cluster yang lain.

Dari hasil cluster 1, terlihat bahwa karakteristik mahasiswa pada cluster 1 didominasi oleh mahasiswa yang berasal dari jurusan Information Technology dan Marketing. Sedangkan, berdasarkan kota asal didominasi oleh mahasiswa yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat, sehingga dapat disimpulkan bahwa rata- rata mahasiswa pada cluster 1 yang berasal dari wilayah kota asal DKI Jakarta dan Jawa Barat mengambil jurusan Information Technology dan Marketing.

2.2.10 Pengujian Hasil Clustering K-Means

Metode pengujian yang digunakan untuk menentukan kriteria penilaian bagus atau tidaknya hasil dari perhitungan *Clustering K-Means* adalah dengan menggunakan metode *Between-Class Variation* (BCV) dan *Within-Class Variation* (WCV) pada iterasi terakhir yang sering disebut dengan rasio. Apabila hasil perhitungan pengujian yang diperoleh besar, maka semakin bagus tingkat kualitas *clustering* tersebut.

BCV merupakan rata-rata dari *centroid*, sedangkan WCV adalah nilai keseluruhan dari jarak minimum yang telah dijumlahkan. Rumus perhitungannya adalah sebagai berikut [20]:

Rumus 2.2 Rumus *Between-Class Variation* (BCV)

$$\text{BCV} = \frac{1}{Nk} \sum_{i}^{k} = 1 \; d(m_i, m_i)$$

Dimana:

k = Jumlah cluster

m_i = Jumlah anggota dari cluster ke-i

i = Nama yang mewakili cluster yang dibentuk

m_i = Jumlah anggota dari cluster ke-i

Rumus 2.3 Rumus *Within-Class Variation* (WCV)

$$\text{WCV} = \sum_{j=i}^{n} \sum_{p \in ci} d(p, m_i)^2$$

Dimana:

p€ci = Jumlah semua data

k = Jumlah cluster

p = Cluster jarak terdekat

m_i = Jumlah anggota dari cluster ke-i

Rasio $= \frac{BCV}{WCV}$

Apabila nilai rasio yang didapat semakin kecil maka semakin bagus pula tingkat hasil dari akurasi cluster [21], kriteria hasil ukuran rasio dapat dilihat pada tabel 2.6.

Table 2.6 Kriteria Pengukuran Rasio

Nilai Rasio	Kriteria
≤ 0,25	Sangat baik
0,25- 0,50	Baik
0,50- 0,75	Kurang baik
0,75– 1,00	Buruk

2.2.11 Metode Receiver Operating Characteristic (ROC)

Tingkat akurasi diukur dengan cara menggunakan metode ROC. Selain mencari nilai akurasi pada metode ini juga dapat dicari nilai sensitivitas dan spesifitas [22], adapun persamaannya dapat dilihat sebagai berikut:

Akurasi $= \frac{Tp+Tn}{Tp+Tn+Fp+Fn}$

Sensifitas $= \frac{Tp}{Tp+Fn}$

Spesifitas $= \frac{Tp}{Tp+Fp}$

Dimana:

Tp = True positif (Nilai kebenaran pada nilai centeroid)

Tn = True negative (Nilai centeroid hasil clustering)

Fp = False positif (Nilai kebeneraan centeroid pada cluser lain)

Fn = False Negative (Nilai kebeneraan centeroid terakhir pada cluser lain)

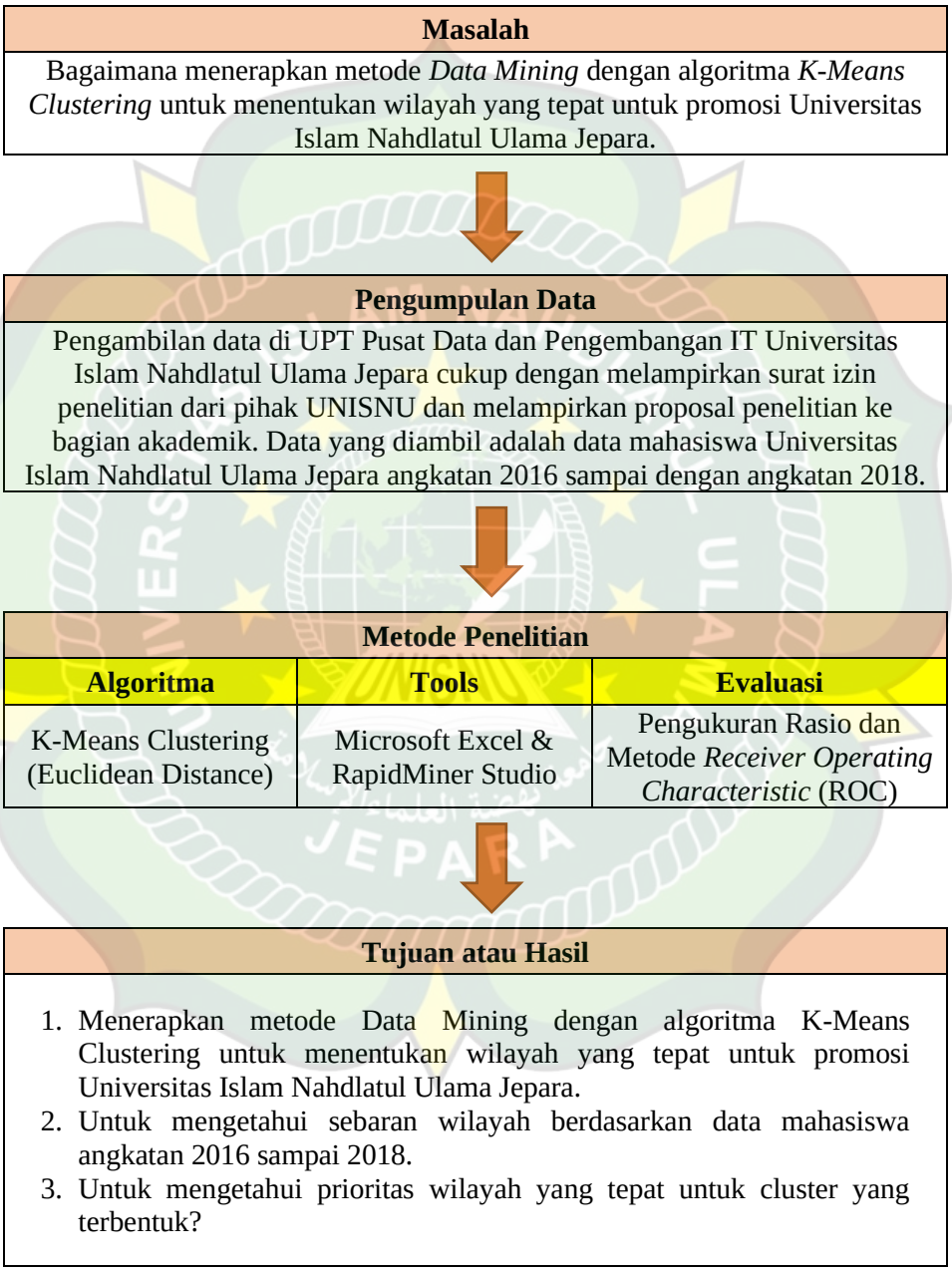
Apabila hasil dari clustering mendekati titik kurva 1,00 maka akurasi yang didapatkan dalam kategori bagus, untuk melihat hasil akurasi masuk kedalam kategori yang mana, perhatikan Tabel di bawah ini.

Table 2.7 Standar Receiver Operating Characteristic (ROC)

Nilai Rasio	Kategori
0,80-1,00	Sangat baik
0,60-0,80	Baik
0,40-0,60	Cukup Baik
0,20-0,40	Kurang Baik
0,00-0,20	Tidak Baik

2.3 Kerangka Pemikiran

Kerangka pemikiran merupakan garis besar dari langkah-langkah penelitian yang sedang dilakukan, kerangka pemikiran dijadikan acuan untuk melakukan tahap-tahap yang sedang dilakukan dalam penelitian.



Gambar 2.7 Kerangka pemikiran