

BAB II

LANDASAN TEORI

2.1. Tinjauan Studi

Dalam melaksanakan penelitian ini, diambil beberapa referensi penelitian sebelumnya diantaranya penelitian yang dilakukan oleh Khoiriya Latifah, dkk (2018) yang berjudul “Analisis dan Penerapan Algoritma C4.5 dalam Data Mining untuk Menunjang Strategi Promosi Prodi Informatika UPGRIS”. Dalam penelitian ini penulis membahas permasalahan tentang bagaimana menentukan strategi promosi perguruan tinggi dengan memanfaatkan data mahasiswa baru yang setiap tahun menumpuk di dalam database sehingga menjadi sebuah informasi yang dapat menunjang promosi yang tepat. Dari masalah yang ada, penulis berencana menggunakan data mahasiswa baru yang berasal dari data saat pendaftaran misalnya nama mahasiswa, jurusan SLTA mahasiswa, nilai UN, asal kota mahasiswa, program studi yang dipilih, IPK, nilai hasil tes masuk perguruan tinggi, asal sekolah, tempat tinggal dan kondisi ekonomi. Kemudian dianalisa dengan menggunakan *data mining* untuk menghasilkan informasi yang berguna dalam alternatif keputusan bagi manajemen perguruan tinggi dalam membuat strategi promosi yang tepat. Metode yang digunakan dalam penelitian ini adalah algoritma C4.5 dengan hasil dari penelitian ini dapat disimpulkan bahwa variabel yang paling tinggi pengaruhnya terhadap registrasi mahasiswa adalah asal sekolah dan jenis kelamin. Interpretasi hasil penelitian mengidentifikasi bahwa variabel yang digunakan sebagai pertimbangan Prodi Informatika dalam melakukan strategi promosi adalah sekolah umum dan dengan rata-rata jurusan IPA berasal dari Semarang, Batang dan Pati. Dari jurusan IPS berjenis kelamin laki-laki berasal dari Batang dan berjenis kelamin perempuan berasal dari Pati. Tingkat akurasi dari model ini adalah 89,33% termasuk dalam klasifikasi yang bagus [1]

Penelitian selanjutnya yang dilakukan oleh Putri Larasati Sibuea dan Andy Sapta, yang berjudul “Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustering”. Permasalahan pada penelitian ini adalah Tingginya tingkat keberhasilan siswa dan rendahnya tingkat kegagalan siswa merupakan cemin kualitas dunia pendidikan. Dunia pendidikan saat ini dituntut untuk memiliki

kemampuan bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Selain sumber daya sarana, prasarana dan manusia. Sistem informasi dapat digunakan untuk meningkatkan kemampuan bersaing. Sistem informasi dapat digunakan untuk menunjang kegiatan operasional sehari-hari sekaligus menunjang kegiatan pengambilan keputusan strategis. Dari masalah yang ada, penelitian ini bertujuan memberikan sebuah informasi dan pengetahuan yang dapat digunakan sebagai bahan pertimbangan pendidik dalam proses pengambilan kebijakan dan keputusan juga sebagai peringatan dini (*early warning*) bagi siswa tertentu yang berdasarkan hasil pengelompokan prestasi rendah yang berpotensi terhadap ketidakhadiran siswa. Metode yang digunakan adalah algoritma *K-Means Clustering*. Atribut data yang digunakan adalah Nama, Ekstrakurikuler, Nilai Tugas, Nilai UTS, Nilai UAS, Jumlah Ketidakhadiran (absensi), dan Nilai Sikap. Dari studi kasus pada 20 siswa dengan perhitungan jarak menggunakan Manhattan distance, Chebyshev distance, dan Euclidean distance menghasilkan akurasi sebesar 67 % [5].

Penelitian dari Koko Handoko yang berjudul "Penerapan Data Mining dalam Meningkatkan Mutu Pembelajaran pada Instansi Perguruan Tinggi Menggunakan Metode K-Means Clustering (Studi Kasus di Program Studi TKJ Akademi Komunitas Solok Selatan)". Adapun permasalahan dari penelitian ini adalah mutu pendidikan di Indonesia yang masih tergolong rendah sehingga diperlukan perbaikan dalam peningkatan mutu pembelajaran pada instansi pendidikan salah satunya perguruan tinggi. Dalam penelitian ini diharapkan dapat memberikan sebuah pengetahuan baru terhadap instansi akademik perguruan tinggi dalam meningkatkan mutu pembelajaran menggunakan data mining. Metode yang digunakan adalah Algoritma K-means Clustering dengan mengelompokkan data yang mirip dan berbeda dari cluster lainnya menggunakan aplikasi Rapidminer 5.3 dengan 4 variabel, yaitu IP mahasiswa, jarak tempuh mahasiswa, jumlah kehadiran dan penghasilan orang tua. Hasil dari penelitian ini adalah algoritma K-means sangat membantu dalam meningkatkan mutu pembelajaran di Akademi Komunitas Solok Selatan dengan data dari IP mahasiswa tergolong sedang, dengan C1 nilainya berkisar 2.50 s/d 3.00. Jarak tempuh mahasiswa tergolong jauh, karena jarak dari C2 tempat tinggal mahasiswa dengan kampus banyak yang 5 KM ke atas. Jumlah kehadiran tergolong sedang dengan jumlah kehadiran dari C3 berjumlah 85 s/d 90

baik yang alfa maupun izin. Penghasilan orang tua tergolong rendah karena penghasilan dari C4 berkisar Rp 0 s/d 1.500.000 [6].

Penelitian-penelitian diatas, tidak sama dengan penelitian ini. Tetapi secara umum aspek-aspek yang dilakukan peneliti terdahulu memberikan dukungan informasi yang diperlukan.

2.2. Tinjauan Pustaka

2.2.1. Definisi Strategi Promosi

A. Pengertian Strategi

Strategi berasal dari bahasa Yunani yaitu *strategosi*, yang berarti jenderal. Oleh karena itu, strategi secara harfiah berarti “seni para jenderal”. Kata ini mengacu pada apa yang merupakan perhatian untuk manajemen puncak organisasi. Secara khusus, strategi adalah “penempatan” misi perusahaan, penempatan sasaran organisasi dengan mengingat kekuatan internal dan eksternal, perumusan kebijakan dan strategi agar mencapai sasaran dan memastikan implementasi tepat, supaya sasaran dan tujuan organisasi akan tercapai [7].

Beberapa tokoh ekonomi menyatakan bahwa setiap perusahaan harus mempunyai strategi untuk menentukan kesuksesan perusahaan itu sendiri. Definisi strategi oleh para pelopor konsep strategi adalah sebagai berikut [7]:

Menurut Itami, Strategi adalah menentukan kerangka kegiatan bisnis perusahaan dan memberikan pedoman untuk kegiatan yang terkoordinasi sehingga perusahaan dapat menyesuaikan dan mempengaruhi lingkungan yang selalu berubah. Sedangkan menurut Andrews, strategi adalah Sasaran dan model kebijakan untuk mencapai tujuan yang ditetapkan, dijelaskan dengan mendefinisikan bisnis yang telah atau harus dijalankan oleh perusahaan [7].

a. Jenis-jenis Strategi

Menurut Griffin, secara umum strategi dibagi dua jenis, dilihat dari lingkungannya, yaitu strategi tingkat perusahaan (*corporate-level strategy*), dan strategi tingkat bisnis (*business-level strategy*). Strategi tingkat perusahaan diterapkan perusahaan untuk persaingan antar perusahaan pada seluruh sektor bisnis. Sedangkan strategi pada level bisnis adalah strategi alternatif yang dilakukan perusahaan untuk persaingan komersial yang dijalankannya pada beberapa jenis bisnis yang diperdagangkan [7].

b. Tipe-tipe strategi

Tipe-tipe strategi disini meliputi stratgi manajemen, strategi investasi, dan strategi bisnis. Adapun penjelasanya adalah sebagai berikut [7]

1. Strategi Manajemen

Manajemen dapat merumuskan arah pengembangan strategi strategi makro seperti, strategi penegembangan produk, strategi akuisisi, strategi penerapan harga, strategi mengenai keuangan, strategi pengembangan pasar, dan sebagainya.

2. Strategi Investasi

Strategi yang berorientasi pada investasi. Misalnya, perusahaan menjalankan strategi pertumbuhan yang agresif atau mengadakan penetrasi pasar, strategi pembangunan kembali divisi baru atau strategi divestasi, strategi pembangunan kembali divisi baru atau strategi divestasi, strategi bertahan, dan sebagainya.

3. Strategi Bisnis

Strategi ini sering disebut strategi bisnis fungsional karena srategi yang berorientasi pada fungsi-fungsi kegiatan manajemen. Misalnya strategi pemasaran, strategi distribusi, strategi organisasi, strategi produksi atau operasional, dan strategi-strategi yang berhubungan dengan keuangan.

B. Pengertian Promosi

Promosi adalah berbagai kegiatan yang dilakukan perusahaan untuk kegiatan tersebut, tujuannya adalah supaya terciptanya proses promosi yang efektif agar tercapai tujuan promosi khususnya dan pemasaran secara keseluruhan yaitu tercapai tingkat penjualan yang diharapkan dan juga memberikan keuntungan bagi masyarakat [7].

Menurut para ahli, promosi adalah [8]:

1. Tjiptono.

Promosi merupakan bentuk komunikasi pemasaran yang berarti aktivitas pemasaran yang menyebarkan informasi, mempengaruhi/membujuk dan mengingatkan pasar atas perusahaan dan produknya agar bersedia menerima, membeli dan loyal pada produk yang ditawarkan perusahaan yang bersangkutan.

2. Swastha dan Irawan.

Promosi adalah aliran informasi satu arah yang digunakan untuk memandu individu atau organisasi mengambil tindakan guna memfasilitasi komunikasi pemasaran.

3. Sistaningrum

Promosi adalah kegiatan perusahaan dalam memengaruhi konsumen aktual ataupun konsumen potensial supaya mau melakukan pembelian terhadap produk yang ditawarkan.

4. Gitosudarmo.

Menurut Gitosudarmo, promosi merupakan kegiatan yang ditujukan untuk mempengaruhi konsumen supaya mengenal produk yang ditawarkan dan kemudian menjadi senang dan membeli produk tersebut.

5. Kotler.

Promosi adalah bagian dari proses strategi pemasaran, cara untuk berkomunikasi dengan pasar dan menggunakan bauran promosi "*promotional mix*".

Dari beberapa pendapat diatas, maka dapat disimpulkan bahwa promosi adalah suatu cara atau kegiatan memperkenalkan produk barang/jasa kepada calon pembeli dengan berbagai cara untuk bisa menarik perhatian pembeli dan hal ini berpengaruh terhadap peningkatan volume penjualan [7].

Adapun tujuan dari promosi itu adalah [7]:

a) Modifikasi tingkah laku

Orang-orang yang melakukan promosi mempunyai beberapa alasan antara lain mencari bantuan, memberikan informasi, memberikan pertolongan, mencari kesenangan, mengemukakan ide dan pendapatan. Sedangkan promosi dari segi lain berusaha merubah tingkah laku dan pendapatan.

b) Memberitahu

Kegiatan promosi dapat didemonstrasikan untuk memberikan informasi tentang produk perusahaan ke pasar yang diharapkan. Pada tahap awal siklus hidup produk, promosi informasi biasanya lebih tepat. Ini adalah masalah yang penting karena orang tidak akan terlibat pada barang atau jasa sebelum mengetahui produk tersebut dan manfaatnya, sehingga bisa memberikan pertimbangan untuk membeli.

c) Membujuk

Promosi yang bersifat membujuk (*persuatif*) terutama diarahkan untuk meningkatkan pembelian. Perusahaan sering kali tidak memerlukan tanggapan sepenuhnya tetapi lebih menginginkan menciptakan kesan positif. Tujuannya adalah untuk memberikan dampak jangka panjang pada perilaku pembeli.

d) Mengingat

Promosi adalah *reminder*, yaitu promosi untuk menjaga merek produk di masyarakat dan perlu dilakukan pada tahap matang dalam daur hidup produk. Ini berarti perusahaan berusaha untuk mempertahankan pembeli yang ada.

Sedangkan tujuan promosi menurut Philip Kotler (1998) adalah:

1. Merangsang para konsumen untuk mencoba suatu barang.
2. Mendekatkan hubungan jarak jauh dengan para pengecer.
3. Menarik para pembeli yang tidak menggunakan merk untuk menggunakan dan memberikan suatu penghargaan kepada mereka yang setia menggunakan merk.
4. Menarik pembeli yang gemar menukar merk dagang.
5. Menawarkan sesuatu sebagian tambahan (bonus) kepada pembeli akan lebih tertarik.

C. Pengertian Strategi Promosi.

Ada beberapa pengertian mengenai strategi promosi, yaitu bahwa strategi promosi adalah kegiatan perusahaan/lembaga untuk menggerakkan penjualan dengan mengarahkan komunikasi-komunikasi yang menyakinkan kepada para pemakai. Ada juga pendapat lain dari Lambert al yang mengatakan bahwa strategi promosi adalah merencanakan penggunaan yang maksimal dari elemen-elemen promosi seperti periklanan, hubungan kepada masyarakat, transaksi pribadi dan marketing penjualan. Dari beberapa pengertian itu dapat diketahui bahwa strategi promosi adalah kegiatan yang disusun secara matang dengan maksud menggoda, mendorong pembeli agar mau atau memiliki keinginan untuk membeli produk perusahaan, sehingga tujuan meningkatkan penjualan dapat dicapai [9].

Terdapat pula dua macam promosi yang ada, yaitu promosi pasif dan promosi aktif. Sesuai dengan judul penelitian ini, promosi pasif digunakan untuk meningkatkan kualitas internal seperti melalui karyawan ataupun dosen UNISNU Jepara, alumni, serta kakak tingkat. Sedangkan promosi aktif adalah melakukan

promosi seperti memasang baliho di beberapa tempat, mencetak brosur dan menyebarkannya, melakukan kunjungan ke SMA maupun SMK dalam rangka mempromosikan Teknik Informatika UNISNU Jepara.

2.2.2. Data Mining

A. Definisi Data Mining.

Data Mining adalah Serangkaian proses untuk menggali nilai tambah yang selama ini tidak dapat diperoleh secara manual dalam bentuk informasi di database [10]. Kata mining berarti usaha mendapatkan sedikit barang berharga dari material dasar. Oleh sebab itu data mining mempunyai akar yang panjang dari bidang ilmu antara lain *machine learning*, kecerdasan buatan (*artificial intelligent*), *statistik*, dan *database* [6]. Informasi yang dihasilkan didapat dari mengekstraksi dan mengenali pola yang penting dari data yang ada di basis data. Data mining digunakan untuk mendapatkan pengetahuan yang ada di basis data yang besar, oleh karena itu sering disebut *Knowledge Discovery Database (KDD)* [10].

Menurut Han, j. Kamber bahwa "*Data Mining* adalah pemilihan atau penggalian pengetahuan dari jumlah data yang banyak". Sedangkan menurut Segall, Guha & Nonis "*Data Mining* disebut penemuan pengetahuan atau penemuan pola tersembunyi dalam data". Dapat disimpulkan bahwa data mining adalah proses menganalisis data dalam jumlah besar dan membuat pola untuk menghasilkan informasi yang berguna [1].

Beberapa teknik yang sering disebut-sebut dalam literatur data mining dalam penerapannya antara lain: *clustering*, *classification*, *association rule mining*, *neural network*, *genetic algorithm*, dan lain-lain [6]. Yang membedakan tanggapan terhadap data mining adalah perkembangan teknik data mining untuk aplikasi database skala besar. Sebelum data mining menjadi populer, teknik ini hanya digunakan untuk data skala kecil. Namun setelah *data mining* semakin populer teknik-teknik tersebut sudah di pakai untuk aplikasi dengan database berskala besar.

B. Fungsi Data Mining

Fungsi-fungsi yang dalam data mining sebagai berikut [10]:

1. Assosiation

Assosiation adalah proses menemukan aturan asosiasi antara kombinasi item pada suatu waktu. *Assosiation* juga disebut *Market Basket Analysis*. Khas dari sebuah problem bisnis adalah menganalisa tabel transaksi penjualan dan mengidentifikasi produk-produk yang selalu dibeli secara bersamaan oleh customer.

2. Sequence

Sequence adalah proses menemukan aturan asosiasi antara kombinasi item pada suatu waktu dan diterapkan lebih dari satu periode.

3. Clustering

Clustering adalah proses untuk mengelompokkan sejumlah data/objek kedalam kelompok data sehingga setiap kelompok berisi data yang mirip.

4. Classification

Classificaion adalah proses penemuan model yang menjelaskan atau membedakan kelas data yang bertujuan memperkirakan kelas dari objek yang labelnya belum diketahui.

5. Regression

Regression adalah proses peemetaan data dalam suatu nilai prediksi. Metode ini hampir mirip metode *Classification*, perbedaannya adalah metode ini tidak bisa menentukan sebuah nilai numerik dan mencari pola yang dijabarkan sebagai pola.

6. Forecasting

Forecasting merupakan Proses memperkirakan nilai prediksi berdasarkan pola dalam sekumpulan data.

7. Solution

Solution merupakan proses menemukan akar masalah dan *problem solving* dari persoalan bisnis yang dihadapi dalam pengambilan keputusan.

C. Proses Tahapan Data Mining

Data mining adalah salah satu rangkaian Knowledge Discovery Database (KDD), yang melibatkan integrasi teknologi dan penemuan ilmiah, interpretasi dan visualisasi pola data. [10]. Beberapa tahap pada proses KDD antara lain [9]:

1. Data selection

Pemilihan data (seleksi data) dari data pokok perlu dilakukan sebelum memulai tahap mengeluarkan informasi dalam KDD. Data yang dipilih akan digunakan dalam proses data mining dan juga dapat dipisahkan dari database operasional.

2. Preprocessing/cleaning

Proses pembersihan data sebelum proses *data mining* dilakukan menjadi inti dari KDD, seperti duplikasi data, kesalahan cetak, data inkonsisten.

3. Transformation

Perubahan pada data yang telah dipilih sehingga sudah sesuai untuk proses *data mining*. Proses ini untuk menciptakan kreasi dan sangat bergantung pada jenis atau pola informasi yang ingin dicari di basis data.

4. Data mining

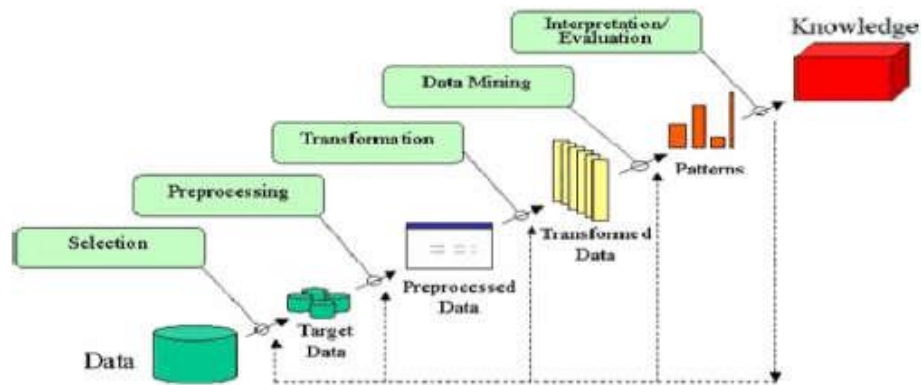
Proses untuk menemukan pola atau informasi pada data yang sudah terpilih dengan menggunakan teknik tertentu. Teknik dalam *data mining* beraneka ragam, penetapan cara atau teknik algoritma yang tepat bergantung penuh pada proses dan tujuan KDD.

5. Interpretation/ evaluation

Proses pada tahap ini meliputi pemeriksaan pola atau informasi data yang dihasilkan apakah sudah benar dan tidak berlawanan dengan fakta atau hipotesis yang sebelumnya telah ada.

6. Knowledge Presentation (Presentasi Pengetahuan)

Pada tahap ini akan terlihat gambaran dan penyajian informasi mengenai teknik atau metode yang digunakan untuk menghasilkan pengetahuan yang didapatkan oleh pengguna.



Gambar 2. 1. Tahapan Knowledge Discovery Database [9].

D. Kategori Data Mining

Data mining dibagi menjadi 2 kategori utama yaitu [10]:

a) Prediktif

Tujuan tugas prediksi adalah untuk memprediksi nilai suatu atribut tertentu berdasarkan nilai atribut lainnya. Atribut prediktif biasanya disebut variabel target atau variabel dependen, dan atribut yang digunakan untuk membuat prediksi disebut variabel independen (*explanatory*).

b) Deskriptif

Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (kolerasi, teritori, cluster, trend, dan anomali) yang memperpendek hubungan pokok pada data.

E. Manfaat Data Mining

Dilihat dari dua sudut pandang, manfaat dari data mining adalah sudut pandang keilmuan dan sudut pandang komersial [10].

- Dari perspektif bisnis, data mining dapat digunakan untuk menghadapi lonjakan volume data, dan teknologi komputasi dapat digunakan untuk menghasilkan informasi yang dibutuhkan, juga merupakan aset yang dapat meningkatkan daya saing suatu organisasi. Contohnya:
 1. Bagaimana cara mengetahui hilangnya pelanggan karena pesaing.
 2. Bagaimana menemukan produk atau produk konsumen dengan karakteristik yang sama.

3. Bagaimana mengidentifikasi produk yang dijual pada saat yang sama dengan produk lainnya .
 4. Bagaimana memprediksi tingkat penjualan.
 5. Bagaimana menilai tingkat resiko saat menentukan volume produksi suatu barang.
 6. Bagaimana memprediksi perilaku bisnis masa depan.
- Dari sudut pandang keilmuan, data mining dapat digunakan untuk mengcapture, menganalisa serta menyimpan data yang bersifat realtime dan sangat besar, misalnya:
 1. *Remote sensor* yang ditempatkan pada suatu satelit.
 2. *Telescope* yang digunakan untuk memindai langit.
 3. Simulasi saintifik yang membangkitkan data dalam ukuran terrabytes.

F. Penerapan Data Mining

1. Analisa Pasar dan Manajemen.

Sumber data yang digunakan antara lain transaksi kartu kredit, kartu keanggotaan klub tertentu, kupon diskon, pengaduan pembeli, dan penelitian gaya hidup masyarakat. Beberapa solusi yang dapat diselesaikan oleh data mining termasuk [10]:

- Menembak target pasar
Data mining dapat mengelompokkan (cluster) model pembeli dan mengklasifikasikan setiap pembeli sesuai dengan karakteristik yang dibutuhkan (seperti kesedihan yang sama, tingkat pendapatan yang sama, kebiasaan membeli dan karakteristik lainnya)..
- Melihat pola beli pemakai dari waktu ke waktu.
Data mining dapat digunakan untuk melihat pola beli dari waktu ke waktu. Sebagai contoh, Setelah seseorang menikah, dia mungkin memutuskan untuk beralih dari satu akun ke akun bersama.
- Cross Market Analysis.
Kita bisa memanfaatkannya untuk melihat keterkaitan antara penjualan satu produk dengan produk lainnya.

- Profil customer.
Data mining dapat melihat profil pelanggan sehingga mereka dapat mengetahui kelompok pelanggan tertentu yang ingin membeli produk apa pun.
 - Identifikasi kebutuhan customer.
Dapat menentukan produk mana yang terbaik untuk setiap kelompok pelanggan dan faktor apa yang dapat menarik pelanggan baru.
 - Menilai loyalitas customer
 - Informasi *Summary*
Dapat digunakan untuk membuat laporan ringkasan multi-dimensi dan memberikan informasi statistik lainnya.
2. Analisa Perusahaan dan Manajemen Resiko
- Perencanaan keuangan dan evaluasi aset
Data mining dapat membantu melakukan analisa dan prediksi *cash flow* serta dapat melakukan *contingent claim analysis* untuk mengevaluasi aset. Selain itu dapat digunakan untuk analisis trend.
 - Perencanaan sumber daya
Dengan melihat ringkasan informasi serta pola pembelanjaan dan pemasukan dari masing-masing *resource*. Maka dapat dimanfaatkan untuk *resource planning*.
 - Persaingan
Data mining dapat membantu memantau pesaing dengan melihat arah pasar mereka. Data mining juga dapat melakukan pengelompokan pelanggan dan dapat memberikan perubahan harga untuk setiap grup.
3. Telekomunikasi
- Data mining melihat jutaan transaksi yang masuk, dan melihat transaksi mana sajakah yang masih harus ditangani secara manual. Tujuannya adalah untuk menambah layanan otomatis.

2.2.3. Clustering

Clustering disebut juga dengan *segmentation*. Metode ini digunakan untuk mengidentifikasi kelompok kasus alami berdasarkan sekumpulan atribut, dan mengelompokkan data dengan atribut serupa [10]. Clustering merupakan Metode data mining yang *unsupervised*, karena tidak ada atribut tunggal yang digunakan untuk memandu pembelajaran, semua atribut masukan dianggap sama [10].

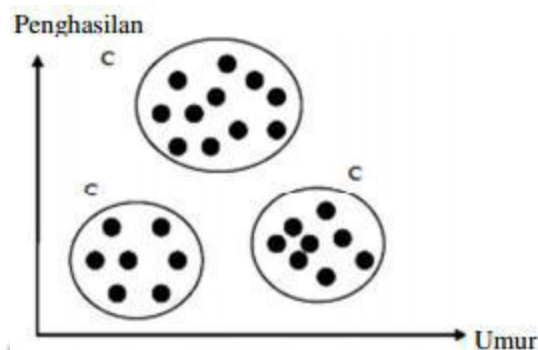
Clustering adalah Teknik mengelompokkan catatan, mengamati atau memperhatikan, dan membentuk suatu kelas objek dengan 18 kesamaan [9]. Menurut Baskoro, *clustering* adalah Salah satu alat data mining yang dirancang untuk mengelompokkan objek data yang mirip dengan satu objek dengan *cluster* yang berbeda [11].

Metode *clustering* yang paling banyak digunakan adalah metode *clustering K-means*. Kerugian utama dari metode ini adalah hasilnya sangat sensitif terhadap pemilihan pusat kluster awal dan solusi lokal untuk mencapai kondisi terbaik. Analisis cluster adalah teknik multivariat yang tujuan utamanya adalah untuk mengklasifikasikan objek sesuai dengan karakteristiknya. Analisis *cluster* mengklasifikasikan objek sehingga setiap objek yang paling mirip dengan objek lain berada dalam cluster yang sama [6].

Terdapat banyak algoritma *Clustering* yang telah digunakan oleh peneliti sebelumnya seperti KMeans, Improved K-Means, Fuzzy C-Means, DBSCAN, K-Medoids (PAM), CLARANS dan Fuzzy Subtractive. Setiap algoritma memiliki kelebihan dan kekurangan masing-masing, namun prinsip algoritma sama, yaitu mengelompokkan data sesuai dengan karakteristik dan mengukur jarak kemiripan antar data dalam satu kelompok [12].

Teknik *cluster* mempunyai dua metode dalam pengelompokkannya yaitu *hierarchical clustering* dan *non-hierarchical clustering*. *hierarchical clustering* merupakan suatu metode pengelompokkan data yang cara kerjanya dengan mengelompokkan dua data atau lebih yang mempunyai kesamaan atau kemiripan, kemudian proses dilanjutkan ke objek lain yang memiliki kedekatan dua, proses ini terus berlangsung hingga *cluster* membentuk semacam tree dimana ada hirarki atau tingkatan yang jelas antar objek dari yang paling mirip hingga yang paling tidak mirip. Namun secara logika semua objek pada akhirnya hanya akan membentuk

sebuah *cluster* [2]. Logika dari konsep *clustering* dapat dilihat pada gambar 2.2 dibawah ini.



Gambar 2. 2. Konsep *Clustering* [2].

2.2.4. Algoritma Clustering K-Means.

Salah satu metode dalam pengelompokan dokumen adalah *K-Means Clustering*. *K-means Clustering* merupakan metode pengelompokan yang paling sederhana, yang membagi data menjadi kelompok-kelompok k berdasarkan centroid masing-masing kelompok. Hanya saja hasil *K-means* yang sangat dipengaruhi oleh parameter k dan inisialisasi sentroid. Namun, metode yang diusulkan akan memodifikasi *K-Means* dalam inisialisasi *centroid*, terutama dalam meningkatkan kinerja pengelompokan dokumen [3].

Algoritma *K-Means* merupakan suatu algoritma yang digunakan untuk menyelesaikan masalah *clustering*. *K-Means* juga merupakan algoritma untuk mengelompokkan atau mengklasifikasi objek berdasarkan atribut/fitur ke dalam k kelompok, dan k merupakan bilangan positif. Tata cara dalam pengelompokan dengan meminimalkan jumlah kuadrat dari jarak antara data dan *cluster centroid* yang bersangkutan. Dengan begitu tujuan pokok dari *K-Means Clustering* adalah untuk mengelompokkan data. Efisiensi serta keakuratan algoritma *K-Means* sangat bergantung kepada *centroid* awal [9].

Algoritma *K-Means* pada dasarnya melakukan 2 proses yakni proses pendeteksian lokasi pusat *cluster* dan proses pencarian anggota dari tiap-tiap cluster. Proses *cluster* dimulai dengan mengidentifikasi data yang akan dikluster, X_{ij} ($i=1, \dots, n; j=1, \dots, m$). Kemudian dihitung jarak antara setiap data dengan setiap pusat *cluster*. Untuk melakukan penghitungan jarak data ke-1 (x_i) pada pusat

cluster ke- k (c_k), diberi nama (c_k), dapat digunakan formula *Euclidean*. Jika jarak ke pusat *cluster* ke- k paling kecil dibandingkan dengan jarak ke pusat *cluster* lain maka data tersebut akan menjadi anggota *cluster* ke- k [2].

Pada dasarnya, saat menentukan pusat *cluster* yang dipilih secara acak dari populasi data, algoritma *K-Means* hanya mendapatkan sebagian dari komponen yang diperoleh, kemudian menjadi pusat *cluster* awal. Kemudian, algoritma *K-Means* akan menguji setiap komponen dalam populasi data dan menandai komponen tersebut sebagai salah satu pusat klaster yang telah ditentukan, tergantung pada jarak minimum antara komponen dan setiap pusat klaster. Selain itu, lokasi pusat *cluster* akan dihitung ulang hingga semua komponen data diklasifikasikan ke dalam setiap *cluster*, dan akan terbentuk *cluster* baru di bagian akhir. [13]. Adapun tahap-tahap melakukan *K-Means Clustering* sebagai berikut [14]:

- a. Pilih nilai k sebagai pusat *cluster* awal secara acak.
- b. Setiap data dalam dataset dibagi kedalam beberapa kelompok k *cluster* antara setiap titik dan pusat *cluster* yang didapatkan berdasarkan pada jarak *Euclidean*. Adapun rumus perhitungan *Euclidean Distance* ditunjukkan pada persamaan berikut :

$$d(x,y) = \| x-y \|_2 = \dots\dots\dots(1)$$

dengan :

d = jarak data ke pusat *cluster*

x = data ke pada atribut ke

y = titik pusat ke, pada atribut

- c. Setiap pusat *cluster* dihitung ulang berdasarkan dari nilai rata-rata dalam *cluster* yang didapatkan.
- d. Langkah a dan b ulangi sampai kelompok tersebut sesuai. Perulangan dapat didefinisikan secara berbeda tergantung pada implementasi, tetapi biasanya pada proses langkah a dan b dapat diulangi apabila kelompok *cluster* masih mengalami perubahan dan proses akan terhenti apabila tidak ada perubahan antar material dalam *cluster*.

2.2.5. Algoritma K-Medoids

K-Medoids atau *Partitioning Around Medoids* (PAM) adalah algoritma *clustering* yang mirip dengan *K-Means*. Perbedaan dari kedua algoritma ini yaitu algoritma *K-Medoids* atau PAM menggunakan objek sebagai perwakilan (*medoid*) sebagai pusat *cluster* untuk setiap *cluster*, sedangkan *K-Means* menggunakan nilai rata-rata (*mean*) sebagai pusat *cluster*. Algoritma *K-Medoids* memiliki kelebihan untuk mengatasi kelemahan pada pada algoritma *K-Means* yang sensitive terhadap noise dan outlier, dimana objek dengan nilai yang besar yang memungkinkan menyimpang pada dari distribusi data. Kelebihan lainnya yaitu hasil proses *clustering* tidak bergantung pada urutan masuk dataset. Langkah-langkah algoritma *K-Medoids*:

1. Inisialisasi pusat *cluster* sebanyak k (jumlah *cluster*)
2. Alokasikan setiap data (objek) ke *cluster* terdekat menggunakan persamaan ukuran jarak *Euclidian Distance* dengan persamaan:

$$d(x,y)=\|x-y\|=\sqrt{\sum_{i=1}^n(x_i-y_i)^2} ; 1,2,3,\dots,n \quad (2)$$
3. Pilih secara acak objek pada masing-masing *cluster* sebagai kandidat *medoid* baru.
4. Hitung jarak setiap objek yang berada pada masing-masing *cluster* dengan kandidat *medoid* baru.
5. Hitung total simpangan (S) dengan menghitung nilai total distance baru – total *distance* lama. Jika $S < 0$, maka tukar objek dengan data *cluster* untuk membentuk sekumpulan k objek baru sebagai *medoid*.
6. Ulangi langkah 3 sampai 5 hingga tidak terjadi perubahan *medoid*, sehingga didapatkan *cluster* beserta anggota *cluster* masing-masing [15]

2.2.6 Davies Bouldin Index

Davies Bouldin Index (DBI) merupakan cara validasi *cluster* yang dibuat oleh D.L. Davies. DBI adalah fungsi rasio dari jumlah distribusi didalam *cluster* untuk pemisahan antar *cluster*. Pengukuran menggunakan DBI bertujuan untuk memaksimalkan jarak inter-*cluster*. Dalam penelitian ini, DBI digunakan untuk melakukan validasi data pada setiap *cluster*. Dengan menggunakan DBI suatu

cluster akan dianggap memiliki skema clustering yang optimal adalah yang memiliki DBI minimal [16].

DBI merupakan salah satu metode evaluasi *cluster* pada suatu metode pengelompokan yang didasarkan pada nilai *kohesi* dan *separasi*. Dalam suatu pengelompokan, *kohesi* didefinisikan sebagai jumlah dari kedekatan data terhadap *centroid* dari *cluster* yang diikuti. Sedangkan *separasi* didasarkan pada jarak antar *centroid* dari *clusternya*. *Sum of square within* (SSW) *cluster* merupakan persamaan yang digunakan untuk mengetahui matrik kohesi dalam sebuah *cluster* ke-*i* yang dirumuskan sebagai berikut [17]:

$$SSWi = \frac{1}{m_i} \cdot \sum_{j=1}^{m_i} d(x_j, c_i)$$

Keterangan:

m_i = jumlah data dalam cluster ke-*i*

c_i = centroid cluster ke-*i*.

Sum of square between cluster (SSB) merupakan persamaan yang digunakan untuk mengetahui *separasi* antar *cluster* yang dihitung menggunakan persamaan :

$$SSB_{ij} = d(c_i, c_j)$$

Setelah nilai *kohesi* dan *separasi* diperoleh, kemudian dilakukan pengukuran *rasio* (R_{ij}) untuk mengetahui nilai perbandingan antara *cluster* ke-*i* dan *cluster* ke-*j*. *Cluster* yang baik adalah *cluster* yang memiliki nilai *kohesi* sekecil mungkin dan *separasi* yang sebesar mungkin. Nilai *rasio* dihitung menggunakan persamaan sebagai berikut :

$$R_{ij} = \frac{SSWi + SSWj}{SSB_{ij}}$$

Nilai *rasio* yang diperoleh tersebut digunakan untuk mencari nilai *Davies Bouldin index* (DBI) dari persamaan berikut :

$$DBI = \frac{1}{K} \sum_i^k \max_{1 \neq j} (R_{ij})$$

Dari persamaan tersebut, k merupakan jumlah *cluster* yang digunakan. Semakin kecil nilai DBI yang diperoleh (non-negatif ≥ 0), maka semakin baik *cluster* yang diperoleh dari pengelompokan *K-means* yang digunakan [18].

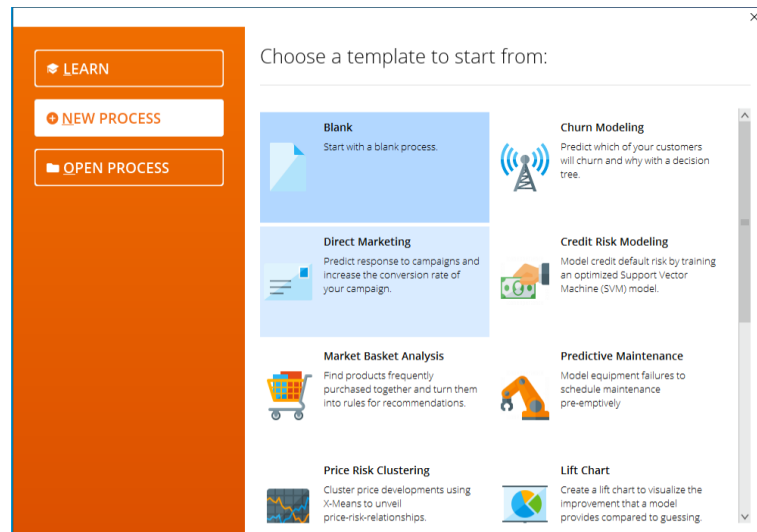
2.2.7. RapidMiner

RapidMiner adalah platform perangkat lunak yang dikembangkan bersama oleh perusahaan ternama yang menyediakan lingkungan terintegrasi untuk persiapan data, pembelajaran mesin, penambahan teks pembelajaran mendalam, dan analisis prediktif. Hal ini digunakan untuk bisnis dan komersial, untuk penelitian, pendidikan, pelatihan, *rapid prototyping*, dan pengembangan aplikasi serta mendukung semua langkah dalam proses pembelajaran mesin termasuk persiapan data, hasil validasi, validasi model, dan optimasi [19].

RapidMiner merupakan salah satu *tool* analisis yang berupa software untuk pengolahan data mining. *RapidMiner* juga sebuah lingkungan (*environment*) untuk *machine learning*, data mining, text mining, dan *predictive analytics*. *Machine learning* adalah suatu algoritma dimana perilaku komputer berevolusi berdasarkan *data empiris*, seperti *sensor* ataupun *database*. *Data mining* merupakan suatu proses merangkai pola-pola dari dataset yang besar dengan mengkombinasikan metode *statistika*, kecerdasan buatan dan *database* [9]. Tujuan dari analisis data adalah untuk mendapatkan informasi berkualitas tinggi dari data yang diproses.

Sejarah *Rapidminer*, sebelumnya dikenal sebagai YALE (Yet Another Learning Environment), mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer dari Unit Kecerdasan Buatan Universitas Teknik Dortmund. Mulai tahun 2006, perkembangan *rapidminer* didorong oleh Rapid-I, sebuah perusahaan yang didirikan oleh Ingo Mierswa dan Ralf Klinkenberg pada tahun yang sama. Pada tahun 2007, nama perangkat lunak itu berubah menjadi *RapidMiner*. Pada tahun 2011, perusahaan melakukan *rebranding* dari Rapid-I menjadi *RapidMiner* [19].

RapidMiner Studio memiliki tampilan antarmuka yang mudah untuk dioperasikan. Ketika dijalankan, *rapidminer studio* menampilkan jendela dialog tiga opsi yaitu : *learn* yang berisi tutorial dasar penggunaan *rapidminer*, *new process* untuk menjalankan proses yang baru dan *open process* untuk membuka proses yang sebelumnya pernah dibuat [19].



Gambar 2. 3. Interface *Rapidminer Studio* [19].

Contoh penggunaan adalah menerapkan sebuah metode pembelajaran ke dataset dan menganalisa hasilnya dan memperoleh informasi tentang data, atau membandingkan beberapa metode dan membandingkan performansinya untuk dipilih. Tools yang dapat pre-processing data yang membuat user dapat berfokus pada algoritma yang digunakan tanpa terlalu memperhatikan detail seperti pembacaan data dari file-file, implementasi algoritma filtering, dan penyediaan kode untuk evaluasi hasil [10].

Tata kerja *Rapidminer* sudah terbukti memberikan 3 tahapan, antara lain [19]:

1. **Transparansi penuh dan tata kelola**
Digunakan untuk pembelajaran mesin, salah satunya memiliki peran penyesuaian yang mudah, yang berarti membuka, memeriksa, dan mengonfigurasi ulang lebih dari 1.500 blok bangunan visual dari algoritme pembelajaran mesin dan operator ilmu data.
2. **Mudah dijelaskan**
Platform otomatis membangun alur kerja analitik visual naratif untuk berkolaborasi dengan pemangku kepentingan bisnis dan menawarkan metode penjelasan model yang kaya..
3. **End to end**
Ini tentang mengonversi data dari sumber mana pun, mengoptimalkan pemilihan dan validasi model *Machine Learning* terbaik.

RapidMiner secara umum akan berperan penting bagi suatu bidang industri, perusahaan, universitas dan organisasi besar untuk meningkatkan mutu dan kualitas.

2.2.8. Metode ROC (*Receiver Operating Characteristic*)

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

Tabel 2. 1. Tabel Metode ROC [20]

Fgh		Kelas Hasil Prediksi	
		Kelas 1	Kelas 2
Kelas Asli (g)	Kelas 1	TP	FN
	Kelas 2	FP	TN
	Total	P	N

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

TP (*True Positive*) = Jumlah data dengan nilai asli positif dan nilai prediksi positif.

FP (*False Positive*) = Jumlah data dengan nilai asli negatif dan nilai prediksi positif.

FN (*False Negative*) = Jumlah data dengan nilai asli positif dan nilai prediksi positif

FP (*False Negative*) = Jumlah data dengan nilai asli negatif dan nilai prediksi negatif

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 di bawah ini [20].

- a. Akurasi bernilai 0.80 – 1.00 = Sangat baik.
- b. Akurasi bernilai 0.6– 0.80 = Baik.
- c. Akurasi bernilai 0.40 – 0.60 = Cukup baik.
- d. Akurasi bernilai 0.20 – 0.40 = Kurang baik.
- e. Akurasi bernilai 0.00 – 0.20 = Tidak baik .

2.2.9. Euclidean Distance

Dalam matematika. *Euclidean distance* digunakan untuk mengukur antara dua titik dalam satu dimensi yang memberikan hasil seperti rumus *Pythagoras*. Persamaan *Euclidean distance* yaitu :

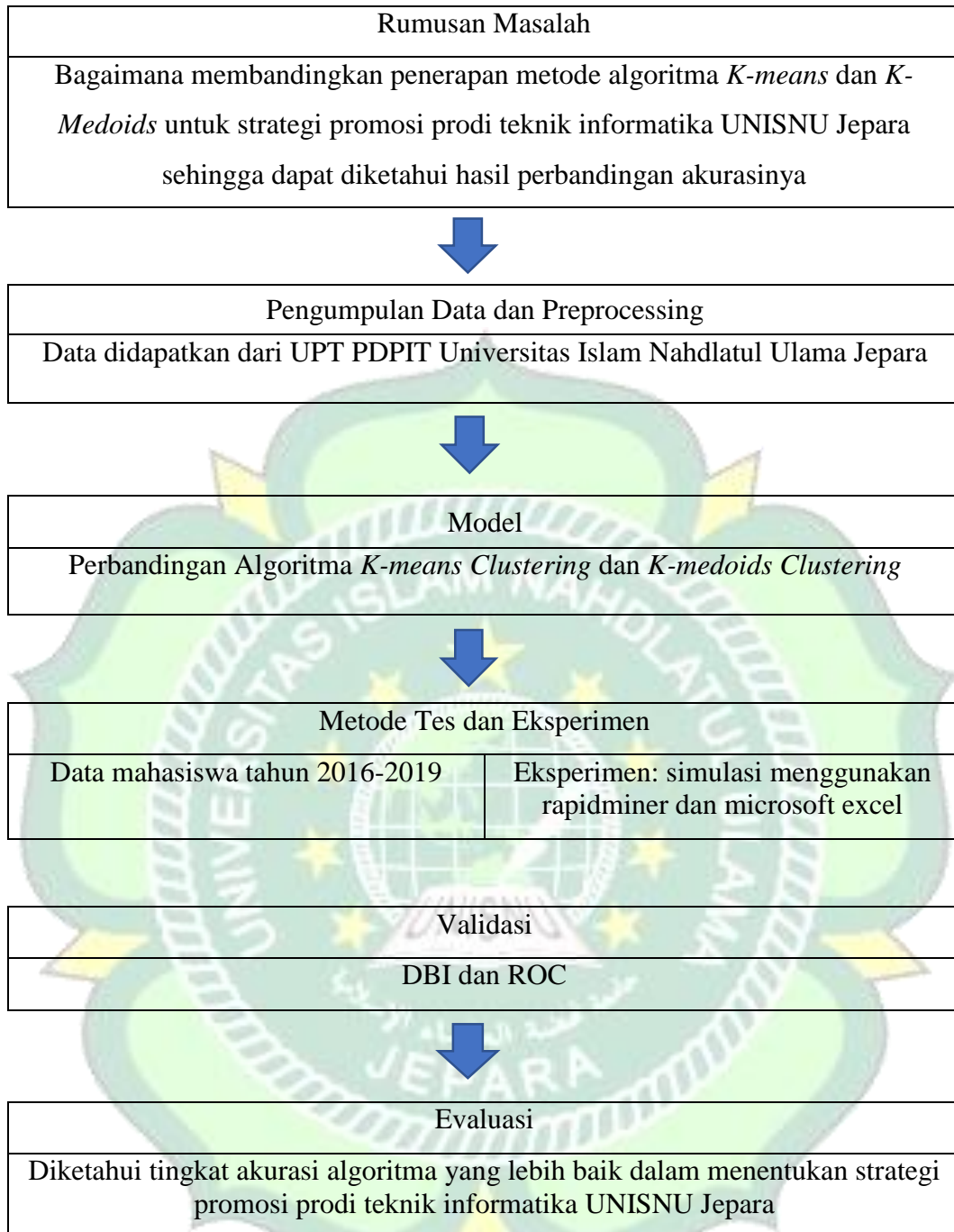
$$D(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

Dari persamaan diatas, $x = (x_1, x_2, x_3, \dots, x_n)$ dan $y = (y_1, y_2, y_3, \dots, y_n)$. *Euclidean distance* diperoleh dari jumlah kuadrat antar titik yang diakar kuadratkan [16].

2.3. Kerangka Pemikiran

Untuk mencapai tujuan penelitian diperlukan perencanaan tahapan kegiatan yang akan dilakukan. Tahapan ini nantinya dijadikan sebagai pedoman pengerjaan penelitian hingga terselesaikannya penelitian ini dengan baik. Berikut adalah proses secara general kerangka tahapan penelitian dapat dilihat pada Gambar 2.5.

Gambar 2.5 menunjukkan diagram penelitian diawali dengan perumusan masalah, seperti yang sudah dijelaskan. Kemudian proses pengumpulan data dan *pre-processing data* dengan data yang didapatkan dari UPT PDPIT UNISNU Jepara. Tahap selanjutnya adalah pemodelan algoritma yang digunakan yaitu peneliti menggunakan algoritma *K-means Clustering* dan *K-Medoids Clustering*. Kemudian tahap selanjutnya adalah metode tes dan eksperimen sesuai skenario yang dirancang serta hasil validasi dan evaluasi untuk penelitian.



Gambar 2. 4. Kerangka Pemikiran [17]