BAB II
Tinjauan Pustaka

2.1. Data Mining

Data mining adalah penemuan informasi baru dengan mencari pola tertentu dari sejumlah data yang sangat besar (Davies dan Beynon, 2004). *Data Mining* juga dapat berarti penambangan sebuah informasi. *Data Mining* juga dikenal dengan istilah *Knowledge Discovery in Database* (KDD) (Santosa, 2007). Segala kegiatan yang meliputi pengumpulan data, pemakaian data, historis untuk menemukan sebuah pola atau hubungan dalam *set* data yang berukuran besar merupakan pengertian dari KDD. Terdapat beberapa karakteristik dari Data *Mining*, yaitu sebagai berikut:

Karakterisitik yang pertama yaitu *Data Mining* berhubungan dengan ditemukannya suatu data atau informasi yang tersembunyi yang tidak diketahui sebelumnya. Lalu yang kedua yaitu *Data Mining* seringkali menggunakan data yang sangat besar dengan tujuan agar hasil yang didapatkan bisa lebih dipercaya. Karakterisitik yang terakhir yaitu *Data Mining* berguna untuk membuat keputusan yang kritis, terutama dalam strategi (Davies dan Beynon, 2004).

Berdasarkan tugas yang dapat dilakukan, *Data Mining* dapat dibagi menjadi beberapa kelompok, yaitu deskripsi, estimasi, prediksi, klasifikasi, pengklasteran dan asosiasi. Secara sederhana peneliti dan analisis mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

Untuk estimasi, variabel targetnya lebih kearah numerik daripada kearah kategori. Nilai dari variabel target sebagai nilai prediksi disediakan oleh model *record* yang dibangun dengan lengkap. Setelah itu, peninjauan berikutnya yaitu estimasi nilai dari variabel target dibuat berdasarkan nilai

variabel prediksi. Antara estimasi, prediksi, dan klasifikasi hampir sama, namun prediksi nilai hasil akan ada di masa yang akan datang. Untuk keadaan yang tepat, beberapa metode dan teknik yang digunkan dalam klasifikasi dan estimasi dapat pula digunakan untuk prediksi.

Klasifikasi hampir sama dengan estimasi dan prediksi. Hanya saja klasifikasi jarang berbentuk numerik. Dalam klasifikasi terdapat target variabel kategori. Kelompok yang terakhir yaitu pengklasteran. Pengklasteran merupakan kegiatan mengelompokkan *record*, mengamati atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. *Cluster* adalah sekumpulan *record* yang meiliki kemiripan satu dengan yang lainnya dan tidak memiliki kemiripan dengan *record-record* dalam *cluster* lain. Perbedaan antara klasifikasi dan pengklasteran yaitu tidak adanya variabel target dalam pengklasteran. Algoritma pengklasteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan(homogen). Menemukan atribut yang muncul merupakan tugas asosiasi dalam *data mining* (Kusrini dan Luthfi, 2009).

Proses dalam KDD secara runtut adalah sebagai berikut (Han dan Pei, 2012):

1. Pembersihan Data (*Data Cleaning*)
   Pembersihan data bertugas untuk menghilangkan *noise* dan data yang tidak konsisten. Pada tahap ini dilakukan penghapusan pada data yang tidak memiliki kelengkapan atribut sesuai yang dibutuhkan.
2. Integrasi Data (*Data Integration*)
   Integrasi data merupakan proses kombinasi beberapa sumber data. Pada tahap ini dilakukan penggabungan data dari berbagai sumber untuk dibentuk penyimpanan data yang koheren.
3. Seleksi Data (*Data Selection*)
   Seleksi data merupakan proses pengambilan data yang berkaitan dengan tugas analisis dari basis data. Pada tahap ini dilakukan teknik perolehan sebuah pengurangan representasi dari data dan

meminimalkan hilangnya informasi data. Hal ini meliputi metode pengurangan atribut dan kompresi data.

4. Transformasi Data (*Data Transformation*)

   Pada tahap ini data diubah dan dikonsolidasikan ke dalam bentuk yang sesuai untuk penambangan (*mining*) dengan melakukan ringkasan atau penggabungan operasi.

5. Penambangan Data (*Data Mining*)

   Data Mining adalah inti pada proses KDD. *Data Mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik tertentu. Pemilihan teknik dan algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

6. Evaluasi Pola (*Pattern Evaluation*)

   Tahap ini merupakan identifikasi kebenaran pola yang merupakan pengetahuan dasar pada langkah-langkah yang diberikan.

7. Representasi Pengetahuan (*Knowledge Presentation*)

   Pada tahap ini penemuan pengetahuan direpresentasikan secara visual kepada pengguna untuk membantu dalam memahami hasil *Data Mining* juga disertakan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

Menurut Pramudiono (2003) ada tiga teknik yang populer didalam *Data Mining*, yaitu Penambangan dengan Aturan Asosiatif, Klasifikasi dan Pengelompokkan. Teknik yang pertama yaitu Penambangan dengan Aturan Asosiatif (*Association Rule Mining*). Penambangan dengan Aturan Asosiatif adalah teknik *mining* untuk menemukan aturan assosiatif antara suatu kombinasi item. Aturan asosiatif digunakan untuk menunjukkan hubungan antara objek data. Aturan asosiatif memiliki dua langkah terpisah: Pertama, mencari minimum *support* yang diterapkan untuk menemukan semua pengulangan *itemset* dalam basis data. Kedua, melakukan pengulangan *itemset* dan menentukan batasan minimum *confidence* yang digunakan untuk membentuk aturannya (Aher dan Lobo, 2012).

Teknik yang kedua yaitu Klasifikasi (*Classification*). Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri dapat berupa aturan jika-maka (*if-then*), berupa pohon keputusan (*decision tree*), jaringan saraf tiruan (*neural network*).

Teknik yang ketiga yaitu Pengelompokan (*Clustering*). *Clustering* melakukan pengelompokan data tanpa berdasarkan kelas data tertentu. *Clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas.

2.2. *Clustering*

Clustering atau klasterisasi merupakan salah satu alat bantu pada *data mining* yang bertujuan mengelompokkan objek-objek ke dalam *cluster-cluster* yang memiliki kemiripan. *Cluster* adalah sekelompok objek-objek data yang memiliki kemiripan satu sama lain dalam *cluster* yang sama dan dimiripkan terhadap objek-objek yang berbeda *cluster.* Kesamaan atau kemiripan objek biasanya diperoleh dari nilai-nilai atribut yang menjelaskan atribut data, objek-objek data biasanya direpresentasikan sebagai sebuah titik dalam ruang mulitidimensi. Objek akan dikelompokkan ke dalam satu atau lebih *cluster* sehingga objek-objek yang berada dalam satu *cluster* akan mempunyai kesamaan yang tinggi antara satu dengan yang lainnya.

Identifikasi daerah yang padat, menemukan pola-pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut-atribut data dapat memanfaatkan penggunaan dari klasterisasi. Beberapa kebutuhan klasterisasi dalam *Data Mining* meliputi skalabilitas, kemampuan untuk menangani atribut yang berbeda, mampu menangani dimensionalitas yang tinggi, menangani data yang mempunyai *noise*, dan dapat diterjemahkan dengan mudah. Tujuan dari data *clustering* yaitu untuk meminimalisasikan *objective function* yang diset dalam

proses *clustering*, yang berusaha meminimalisasikan variasi di dalam suatu *cluster* dan memaksimalkan variasi antar *cluster* (Nango, 2012).

Pada *clustering* terdapat dua pendekatan umum, yaitu pendekatan partisi dan pendekatan hirarki. Pendekatan partisi merupakan pengelompokkan data dari satu kelompok besar kemudian dibagi menjadi beberapa kelompok yang lebih kecil. *K-Means Clustering* merupakan salah satu contoh pendekatan ketika menggunakan metode *clustering*. Pendekatan hirarki atau yang biasa disebut dengan *Hierarchical Clustering* adalah mengelompokkan data dengan menggabungkan masing-masing *record* atau individu pada data menjadi *cluster-cluster*. *Agglomerative Hierarchical Clustering* merupakan salah satu contoh metode *clustering* dengan pendekatan hirarki (Yusuf dan Tjandrasa, 2013). Teknik *clustering* memiliki penggunaan yang luas dan saat ini memiliki kecenderungan yang semakin meningkat seiring dengan jumlah data yang terus berkembang (Sharma dkk, 2012).

Clustering menggunkan metode *K-Means* secara umum dilakukan dengan algoritma sebagai berikut:

1. Tentukan jumlah *cluster*
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat *cluster (centroid)* menggunakan *mean* untuk masing-masing kelompok
4. Alokasikan masing-masing data ke *centroid* terdekat
5. Kembali ke langkah 3, jika masih ada data yang berpindah *cluster* atau jika nilai *centroid* diatas nilai ambang, atau jika nilai pada fungsi obyektif yang digunakan masih diatas ambang

Metode ini mencoba menemukan pusat dari kelompok dalam data sebanyak iterasi perbaikan yang dilakukan. Metode ini berusaha membagi data kelompok sehingga data yang berkarakteristik sama dimasukan ke dalam satu kelompok sementara data yang berkarakteristik berbeda dimasukkan dalam kelompok lain. Pada langkah 3, lokasi *centroid* setiap kelompok diambil dari rata-rata semua nilai data pada setiap fiturnya. Jika M menyatakan jumlah, I menyatakan

fitur/variabel/atribut ke=i dan p menyatakan dimensi dari data, untuk menghitung *centroid* fitur ke I digunakan formula : $C_{i=\frac{1}{M}\sum_{j}^{M}=1xj}$ …….....(2.1)

Jarak antara data dan *centroid* diukur dengan beberapa cara, diantaranya :

Euclidean: $D(X_2, X_1) = ||X_2 - X_1|| = \sqrt{\sum_{i=1}^{p} |X_{2j} - X_{1j}|^2}$ …….....(2.2)

D adalah jarak antara data X2 dan X1, dan |.| adalah nilai mutlak.

Pengukuran jarak pada ruang jarak *Manhattan* menggunakan formula :

$D(X_2, X_1) = ||X_2 - X_1||_2 = \sum_{j=1}^{p} |X_{2j} - X_{1j}|$ …….....(2.3)

Pengukuran jarak pada ruang jarak Minkowsky :

$D(X_2, X_1) = ||X_2 - X_1||\lambda = \lambda \sqrt{\sum_{i=1}^{p} |X_{2j} - X_{1j}|^{\lambda}}$ …….....(2.4)

Euclidean dan Manhattan merupakan cara yang paling banyak digunakan. Euclidean menjadi pilihan jika ingin dicari jarak terpendek antara dua titik, sedangkan Manhattan memberikan jarak terjauh antara dua titik.

Pada langkah 4, pengalokasian data ke dalam masing-masing kelompok pada *K-Means* didasarkan pada perbandingan jarak antara data dengan *centroid* setiap kelompok yang ada. Pengalokasian dirumuskan sebagai berikut :

$a_{ij} = \{_{0}^{1} d = \min\{D(X_i, C_1)\}$ …….....(2.5)

a_{ij} adalah nilai keanggotaan titik X_i ke *centroid* C_1 , d adalah jarak terpendek dari data X_i ke k kelompok setelah dibandingkan, dan C1 adalah *centroid* ke-1.

Menurut Han dan Pei (2006) secara umum metode pada *clustering* dapat digolongkan ke dalam beberapa metode yaitu: Metode Partisi, Metode Hirarki, Metode Berbasis Kerapatan, Metode Berbasis Grid, dan Metode Berbasis Model.

Langkah kerja metode partisi, yaitu apabila terdapat basis data sejumlah n objek atau data tupelo, selanjutnya data di partisi menjadi k partisi dari data, dimana setiap partisi mewakili sebuah *cluster* dan k ≤ n. Adapun syarat yang harus terpenuhi sebagai berikut: (1) setiap kelompok harus berisi setidaknya satu objek, dan (2) setiap objek harus memiliki tepat satu kelompok. Awalnya basis data dipartisi menjadi k partisi. Kemudian menggunakan teknik relokasi berulang, mencoba

untuk memperbaiki partisi dengan memindahkan dari satu kelompok ke kelompok lain. Kriteria umum dari partisi yang baik adalah bahwa objek dalam satu *cluster* memiliki kemiripan yang sangat dekat, sedangkan objek dalam *cluster* yang berbeda memiliki kemiripan yang jauh berbeda. Pencapaian optimalitas global dalam pengelompokan berbasis partisi akan memerlukan penghitungan lengkap dari semua partisi yang memungkinkan. Sebaliknya, sebagian besar aplikasi mengadopsi salah satu dari beberapa metode heuristik yang popular, seperti (1) algoritma *K-Means*, dimana setiap segmen diwakili oleh nilai rata-rata dari objek dalam *cluster*, dan (2) algoritma *K-Medoids*, dimana setiap segmen diwakili oleh salah satu objek yang terletak didekat *centroid*. Metode pengelompokan heuristik ini bekerja dengan baik untuk menemukan *cluster* berbentuk bola kecil untuk basis data yang berukuran sedang.

Metode hirarki menciptakan dekomposisi hirarki dari himpunan objek data yang diberikan. Sebuah metode hirarki dapat diklasifikasikan sebagai salah satu *agglomerative* atau memecah belah, berdasarkan cara dekomposisi hirarki terbentuk. Pendekatan *agglomerative* memiliki dua cara pendekatan, yaitu *bottom–up* dan *top-down*. Pendekatan *bottom-up* berlangsung seperti berikut, awalnya setiap objek membentuk kelompok tersendiri. Berturut-turut menggabungkan objek atau kelompok yang dekat satu sama lain, sampai semua kelompok digabung menjadi satu (tingkat teratas dari hirarki), atau sampai terjadinya kondisi pemutusan hubungan. Sedangkan pendekatan *top down*, dimulai dengan semua objek dalam *cluster* yang sama dibagi menjadi kelompok yang lebih kecil, sampai akhirnya setiap objek dalam satu *cluster* atau sampai terjadi kondisi pemutusan hubungan. Metode hirarki memuat fakta bahwa setelah langkah penggabungan atau *split* dilakukan, proses memecah belah tidak dapat dibatalkan. Ada dua pendekatan untuk meningkatkan kualitas pengelompokan hirarki: (1) melakukan analisis yang cermat terhadap objek "*linkage*" pada setiap partisi hirarki, seperti di *Chameleon*, atau (2) mengintegrasikan aglomerasi hirarki dan pendekatan-pendekatan lain dengan terlebih dahulu menggunakan algoritma *agglomerative* hirarki objek kelompok ke dalam *microclusters*, dan kemudian melakukan *macroclustering* pada *microclusters* menggunakan metode pengelompokan lain seperti relokasi berulang. Salah satu algoritma yang tergolong kedalam metode hirarki yaitu BIRCH

(*Balanced Iterative Reducing and Clustering Using Hierarchies*). BIRCH merupakan salah satu algoritma pengelompokan hirarki yang terintegrasi. BIRCH memperkenalkan dua konsep, *clustering feature* dan *clustering feature tree* (CF *tree*), yang mana digunakan untuk menggambarkan ringkasan *cluster*.

2.3. Algoritma *K-Means*

K-Means merupakan algoritma *clustering* yang berulang-ulang. Algoritma *K-Means* dimulai dengan pemilihan secara acak *cluster* yang ingin dibentuk. Setelah memilih acak *cluster* yang akan dibentuk, langkah selanjutnya yaitu menetapkan nilai-nilai *cluster* secara *random,* untuk sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan *centroid, mean* atau "*means".* Setelah itu menghitung jarak setiap data yang ada terhadap masing-masing *centroid* menggunakan rumus Euclidian sampai ditemukan jarak yang paling dekat dari setiap data dengan *centroid.* Langkah selanjutnya yaitu mengklasifikasikan setiap data berdasarkan kedekatannya dengan *centroid.* Klasifikasi ini dilakukan hingga nilai *centroid* satbil atau tidak berubah (Sharma dkk, 2012). Algoritma *K-Means* pada dasarnya melakukan dua proses, yang pertama yaitu pendeteksian lokasi pusat tiap *cluster* dan proses pencarian anggota dari tiap-tiap *cluster*. *K-Means* adalah teknik sederhana untuk analisis *clustering.* Tujuannya adalah untuk menemukan divisi terbaik entitas data ke dalam kelompok *cluster*, sehingga total jarak antara anggota kelompok dan *entroid* sesuai, terlepas dari kelompok diminimalkan (Sharma dkk, 2012).

Efisiensi keaslian algoritma *K-Means* sangat bergantung pada titik pusat *cluster* (*centroid*) awal (Yedla dkk, 2010). Langkah kerja dari algoritma *K-Means* adalah sebagai berikut :

1. Menanyakan kepada pengguna berapa banyak k *cluster* dataset yang akan dipartisi.
2. Menetapkan secara acak k *record* yang menjadi lokasi pusat *cluster* awal.

3. Setiap *record* dicari *centroid cluster* terdekatnya. Artinya setiap *centroid cluster* "memiliki" subset dari *record*, sehingga merepresentasikan sebuah partisi dari dataset. Didapatkan k *cluster*, C1, C2, …, Ck.
4. Setiap k *cluster* dicari centroidnya dan memperbarui lokasi setiap pusat *cluster* untuk nilai centroid *baru*.
5. Ulangi langkah 3 sampai 5, sampai terjadi konvergensi atau terjadi penghentian.

Algoritma berakhir ketika titik pusat *cluster* tidak lagi berubah. Dengan kata lain, algoritma berakhir ketika dari seluruh *cluster* C1, C2, …, Ck, semua *record* yang dimiliki oleh masing -masing pusat *cluster* tetap dalam *cluster* itu.

2.4. IKM (Industri Kecil Menengah)

Definisi UMKM menurut Undang-Undang No. 20 Tahun 2008 tentang Usaha Mikro, Kecil dan Menengah Bab 1 Pasal 1: "Usaha mikro adalah usaha produktif milik orang perorangan dan tau badan usaha perorangan yang memenuhi kriteria usaha mikro. Usaha kecil adalah usaha produktif yang berdiri sendiri, yang dilakukan oleh perseorangan atau badan usaha bukan merupakan anak cabang perusahaan yang dimiliki, dikuasai, atau menjadi bagian baik langsung maupun tidak langsung dari usaha menengah atau besar yang memenuhi kriteria usaha kecil. Usaha menengah adalah usaha ekonomi produktif yang berdiri sendiri, yang dilakukan oleh orang perorangan atau badan usaha yang bukan merupakan anak perusahaan atau cabang perusahaan yang dimiliki, dikuasai, atau menjadi bagian baik langsung maupun tidak langsung dengan Usaha kecil atau Usaha besar dengan jumlah kekayaan bersih atau hasil penjualan tahunan.

Definisi UMKM menurut Kementrian Koperasi dan UMKM yaitu: Usaha Kecil (UK), termasuk Usaha Mikro (UMI) adalah entitas usaha yang memiliki kekayaan bersih paling banyak Rp. 200.000.000, tidak termasuk tanah dan bangunan tempat usaha dan memiliki penjualan tahunan paling banyak Rp. 100.000.000. Sementara itu, Usaha Menengah (UM) merupakan entitas usaha milik warga negara Indonesia

yang memiliki kekayaan bersih lebih besar dari Rp. 200.000.000 s.d. Rp. 10.000.000.000 tidak termasuk tanah dan bangunan.

Secara garis besar, UMKM dan IKM terlihat sama. Namun yang menjadi pembeda dari kedua istilah tersebut yaitu di dalam IKM selalu terdapat proses produksi sendiri oleh masing-masing usaha. Sedangakan pada UMKM, tidak semua usahanya melakukan proses produksi sendiri, hanya berjualan saja. IKM adalah Industri Kecil Menengah. Menurut Peraturan Kementrian Perindustrian No. 64 tahun 2016 industri kecil adalah industri yang memiliki karyawan maksimal 19 orang, memiliki nilai investasi kurang dari 1 miliar rupiah, tidak termasuk tanah dan bangunan tempat usaha. Sedangkan yang dimaksud dengan industri menengah adalah industri yang memiliki karyawan maksimal 19 orang dan nilai investasi minimal 1 miliar rupiah atau memiliki karyawan minimal 20 orang dan nilai investasi maksimal 15 miliar rupiah.

2.6 Penelitian Terdahulu

Menurut Retnaningsih (2016) dalam skripsinya yang berjudul "Pengklasifikasian Data Sekolah Pengguna Internet Pendidikan Menggunakan Teknik *Clustering* Dengan Algoritma *K-Means* Pada Studi Kasus PT.Telkom Surabaya Universitas Nusantara PGRI Kediri" mampu menjadi referensi saya dalam menyusun strategi pemecahan masalah klasifikasi dan *clustering* yang dialami oleh IKM *Center* Kabupaten Malang. Teknik *clustering* sangat efektif digunakan untuk mengklasifikasikan data numerik dengan jumlah yang sangat besar. Kesamaan penelitian ini dengan penelitian yang terdahulu yaitu sama-sama menggunakan teknik *clustering* dengan metode *K-Means* dan juga teknik klasifikasi dengan metode *Decision Tree.* Perbedaannya yaitu objek penelitian dan data yang digunakan. Data yang digunakan pada penelitian terdahulu yaitu pengguna internet pendidikan di PT. Telkom Surabaya.

Menurut Defiyanti (2017) dalam jurnalnya yang berjudul "Integrasi Metode *Clustering* dan Klasifikasi untuk Data Numerik Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang" mampu menjadi referensi saya untuk

melakukan klasifikasi dan *clustering* dengan menggunakan *tools* Weka 3.8. Selain itu, penelitian terdahulu juga menggunakan metode *Data Mining.* Persamaan antara penelitian terdahulu dengan penelitian ini yaitu sama-sama dimulai dengan proses *Data Mining*, selanjutnya dilakukan pemodelan klasifikasi dan *clustering* dengan menggunakan *tools* Weka 3.8. Perbedaan penelitian terdahulu dengan penelitian ini yaitu data yang digunakan. Penelitian terdahulu menggunakan dataset *Diabetic Retinopathy Debrecen Data Set*. *Diabetic Retinopathy Debrecen Data Set* merupakan data citra digital yang digunakan untuk mendeteksi *Diabetic Retinopathy*.