BAB II
DASAR TEORI

2.1 Penelitian Terkait

Adapun penelitian-penelitian terdahulu yang terkait dengan tugas akhir ini sebagai berikut:

2.1.1 Data Transformation pada Data Mining

Penelitian ini dilakukan oleh Junaedi dkk pada tahun 2011 dengan judul "*Data Transformation* pada *Data Mining*". Pada penelitian Junaedi dkk. membahas beberapa metode yang terdapat pada *data transformation* dan melakukan analisis metode yang dimiliki untuk masing-masing metode tersebut. Metode atau operasi transformasi yang dibahas di sini, yaitu *Smoothing* (*Binning*, *Clustering*, dan *Regression*), *Generalization* (*Histogram Analysis*, *Entropy-Based Discretization*, dan *Segmentation by Natural Partitioning*), *Normalization* (*Min-max Normalization*, *Z-Score Normalization*, dan *Normalization by Decimal Scaling*), *Aggregation* (*Roll-up*), dan *Attribute Construction*.

2.1.2 *Cross-company Customer Churn Prediction (CCCP) in Telecommunication: A Comparison of Data Transformation Methods*

Penelitian ini dilakukan oleh Amin dkk pada tahun 2018 dengan judul "*Cross-Company Customer Churn Prediction* (CCCP) *in Telecommunication: A Comparison of Data Transformation Methods*". Tujuan dari penelitian ini untuk memprediksi persentase pelanggan dengan menggunakan data dari perusahaan lain sebagai sumber. Pada penelitian ini, mencari metode transformasi data pada kinerja model CCCP yang efektif. Metode yang diujicobakan adalah *log*, *z-score*, *rank*, dan *box-cox*. Hasilnya menunjukkan bahwa sebagaian metode transformasi data dapat meningkatkan kinerja CCCP secara signifikan. Namun, metode transformasi data *z-score* tidak dapat mencapai hasil yang lebih baik dibandingkan dengan metode transformasi data lainnya dalam penelitian ini (Amin et al., 2019).

2.1.3 Posisi Penelitian

Dengan referensi penelitian terdahulu, informasi dari penelitian tersebut akan menjadi pelajaran dalam menggunakan metode *data transformation*. Data yang menjadi objek penelitian sudah mengalami proses *data cleansing*. Pada tugas akhir ini, data tersebut akan

diulas bagaimana penerapan *data transformation* terjadi pada Sistem Informasi Manajemen Rumah Sakit. Pada hasil tugas akhir ini, diharapkan data yang sudah ditransformasikan siap untuk dilakukan proses *mining*.

Tabel 2.1 Tabel komparasi penelitian

Judul Penelitian	Metode	Tujuan
Data Transformation pada Data Mining	*Smoothing*, *Generalization*, *Normalization*, A*ttribute construction*, dan Algoritma Tambahan	Menganalisis Metode *Data Transformation* pada Aplikasi Weka
Cross-Company Customer Churn Prediction (CCCP) *in Telecommunication: A Comparison of Data Transformation Methods*	*Log, Z-score, Rank, Box-cox,* dan *Min-max*	Mencari Metode *Data Transformation* pada kinerja model CCCP
Penerapan Data Transformation pada *Database* Sistem Informasi Manajemen Rumah Sakit	*Smoothing, Attrubute Construction, Discretization, Normalization,* dan *Aggregation*	Menerapkan Metode *Data Transformation* pada *Database* SIMRS

2.2 *Smoothing*

Smoothing adalah salah satu transformasi data yang menggunakan hubungan nilai-nilai tetangga dengan mengkategorikan data ke dalam inteval berdasarkan karakteristik (Dua & Chowriappa, 2012). Hal ini bertujuan untuk menghaluskan data dan mengurangi kontras pada data. *Smoothing* dapat diimplementasikan jika data tersebut mengandung *noise*/nilai yang tidak valid. Dalam mengatasi hal ini harus dilakukan *smoothing* (dengan memperhatikan nilai-nilai tetangga). Berikut teknik atau metode untuk *smoothing*:

2.2.1 *Binning*

Metode *Binning* memiliki kecenderungan untuk menghaluskan data yang diurutkan dengan memeriksa nilai tetangganya, yaitu data di sekitarnya (Baskar, Arockiam, & Charles, 2013). Berikut adalah langkah-langkah metode *binning*:

a. Data diurutkan terlebih dahulu dari yang terkecil hingga terbesar.
b. Data yang sudah diurutkan kemudian dipartisi ke dalam beberapa *bin*. Terdapat dua cara dalam membagi partisi ke dalam *bin*, yaitu *equal-width* (*distance*) *partitioning* dan *equal-depth* (*frequency*) *partitioning*.
c. Ada tiga macam cara dalam melakukan *smoothing* pada setiap *bin*, yaitu *smoothing by bin-means, smoothing by bin-medians*, dan *smoothing by bin-boundaries*.

2.2.2 *Clustering*

Clustering adalah proses membagi partisi atau mengelompokkan serangkaian pola yang diberikan ke dalam *cluster* yang terpisah (Alsabti, 1997). *Clustering* berguna untuk menyingkirkan *outliers*. Pada metode *clustering* dapat menggunakan algoritma *k-means* yang merupakan kategori metode *partitioning.* Algoritma ini dapat digunakan jika ukuran *database* tidak terlalu besar. Menurut Junaedi dkk. algoritma *k-means* ini berdasarkan pada nilai tengah dari objek sudah dideklarasi yang ada dalam tiap *cluster*. *K-means* meminta data masukan dengan parameter *k* dan membagi partisi menjadi satu set *n* objek ke dalam *k cluster* kemudian akan menghasilkan tingkat kemiripan yang besar antar objek dalam satu kelas yang sama (*intra-class similarity*) dan tingkat kemiripan yang kecil antar objek dalam kelas yang berbeda (*inter-class similarity*). Kemiripan *cluster* ini diukur dengan cara menghitung nilai tengah dari objek yang ada di dalam tiap *cluster*.

2.3 *Attribute Feature Construction*

A*ttribute/feature construction* adalah proses untuk menemukan informasi yang hilang dengan membuat atribut/fitur baru yang dibentuk dari atribut yang sudah ada untuk membantu meningkatkan ketelitian/ketepatan (Motoda & Liu, 2002). Contohnya, jika ingin menambahkan atribut umur berdasarkan atribut tanggal kelahiran atau atribut kategori umur dari umur digantikan dengan anak-anak, remaja, dan dewasa.

2.4 *Normalization*

Normalization atau normalisasi adalah proses pengelompokan atribut ke dalam hubungan yang terstruktur dengan baik dan bebas dari anomali (Lee, 1995). *Normalization* digunakan untuk mentransformasi sebuah atribut numerik diskalakan dalam *range* yang lebih kecil seperti dari *range* -1.0 sampai dengan 1.0. Ada beberapa metode/teknik yang diterapkan untuk normalisasi data, di antaranya:

2.4.1 *Min-Max Normalization*

Min-max normalization memetakan sebuah nilai *v* dari atribut *A* menjadi *v*' ke dalam range [new_{min_A}, new_{max_A}] berdasarkan persamaan (2.4) (Junaedi et al., 2011).

$$v' = \frac{v - min_A}{max_A - min_A} . (new_{max_A} - new_{min_A}) + new_{minA} \quad (2.1)$$

2.4.2 *Z-Score Normalization*

Bisa disebut juga dengan *zero-mean normalization*, dimana nilai dari sebuah atribut *A* dinormalisasi berdasarkan nilai rata-rata dan standar deviasi dari atribut *A* (Junaedi et al., 2011). Sebuah nilai *v* dari atribut *A* dinormalisasi menjadi *v*' dengan persamaan (2.2), dimana $\overline{A}$ dan σ_A adalah nilai rata-rata dari standar deviasi dari atribut *A.*

$$v' = \frac{v - \overline{A}}{\sigma_A} \quad (2.2)$$

2.5 *Aggregation*

Aggregation adalah operasi yang menerapkan proses meringkas/*summary* pada data (Han et al., 2012). Biasanya operasi diimplementasikan pada data numerik. Misalnya pada data penjualan harian dikumpulkan sehingga dapat menghitung jumlah total atau rata-rata perbulan dan pertahun. Langkah ini biasanya digunakan dengan memanfaatkan operator *data cube* untuk analisis data pada beberapa level abstraksi.

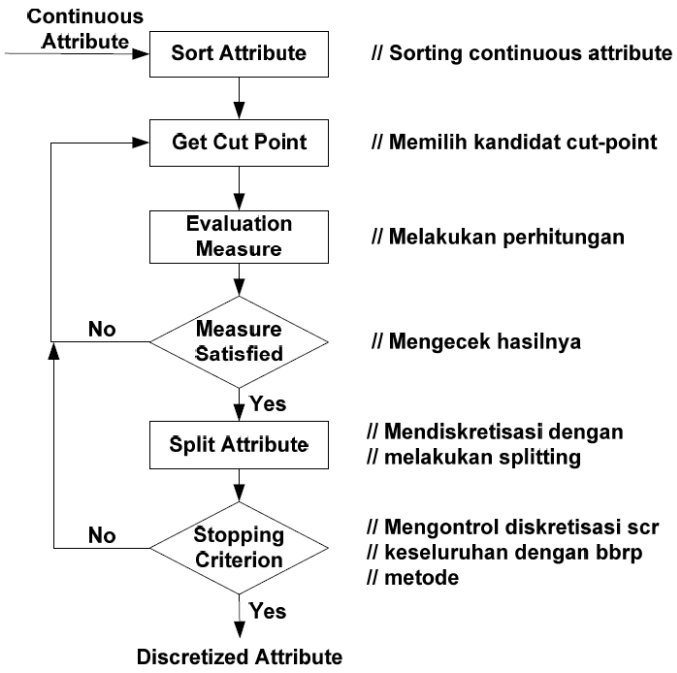
2.6 *Discretization*

Discretization atau *generalization* adalah proses mengubah data level rendah (*low-level data*) diganti dengan dengan label interval/konsep yang lebih tinggi (Han et al., 2012). Teknik diskretisasi digunakan untuk mengurangi sekumpulan nilai yang terkandung dalam atribut *continuous* dengan cara membagi *range* dari atribut ke dalam interval.

Proses diskretisasi terdapat empat tahapan, yaitu:

a. *Sorting*. mengurutkan nilai atribut yanag bersifat *continuous* yang akan didiskretisasi.
b. Menentukan *"cut-point"* dengan cara menggunakan fungsi seperti *binning* dan pengukuran *entropy*.
c. *Splitting*. Setelah melakukan *cut-point,* dilakukannya evaluasi dan memilih satu yang terbaik dan lakukan pembagian/*splitting range* nilai atribut *continuous* ke dalam dua partisi. Diskretisasi terus dilanjutkan sampai kondisi berhenti tercapai.
d. *Stopping criterion*. Tahap ini sebagai pemberhentian proses diskretisasi.

Pada *discretization* terdapat dua metode yang digunakan untuk proses diskretisasi pada atribut kontinu, yaitu *binning* dan *clustering*. Kedua metode tersebut, sebelumnya sudah dijelaskan pada teknik *data smoothing*.



Gambar 2.1 Proses disketisasi

Sumber: Junaedi dkk. (2011)

2.7 Kakas Analisis

Alat yang digunakan untuk membantu dalam menjalankan metode yang telah disebutkan di atas adalah RapidMiner. RapidMiner adalah sistem yang mendukung desain dan dokumentasi dari keseluruhan proses penambangan data (Markus Hofmann Ralf Klinkenberg, 2009). RapinMiner menyediakan lingkungan yang terintegrasi dalam hal persiapan data, pembelajaran mesin, penambangan teks, dan analisis secara prediktif. Alat ini sangat cocok untuk bisnis, penelitian, pendidikan, pelatihan, *rapid prototyping*, dan pengembangan aplikasi serta mendukung semua langkah dalam proses pembelajaran mesin termasuk persiapan data, hasil visualisasi, validasi model, dan optimasi.