BAB II

TINJAUAN PUSTAKA

II.1 Penelitian Terkait

Dalam penelitian terdahulu, penulis mengambil beberapa referensi yang berkaitan dengan penerapan metode algoritma k-means sebagai penelitian dalam pengelompokkan data diantaranya yaitu penelitian yang dilakukan oleh Suryani Retno Widyaningrum, 2016 "Implementasi *Data Mining* Untuk Pengelompokkan Data Siswa Menggunakan Algoritma K-Means *Clustering* (Studi Kasus : SMKN 1 Kediri)" yang berkesimpulan bahwa telah terbentuknya 4 kelompok yang menunjukkan bagaimana nilai siswa saat berada di sekolah yang nantinya berguna untuk pihak sekolah dalam memantau perkembangan belajar siswa.

Penelitian yang dilakukan oleh Widya Safira Azis dan Dedy Atmajaya (2016) dalam penelitiannya yang berjudul "Pengelompokkan Minat Baca Mahasiswa Menggunakan Metode K-Means", penulis dalam penelitan ini menjelaskan bahwa salah satu cara untuk mengelola data koleksi buku yang ada di perpustakaan yaitu dengan cara data mining dengan menggunakan algoritma k-means. Data dikelompokkan menjadi 3 *cluster* yang nantinya hasil akhirnya berupa centroid yang mempunyai nilai terbesar merupakan *cluster* yang akan dijadikan rekomendasi dalam penambahan koleksi buku.

Penelitian yang dilakukan oleh Fina Nasari dan Surya Darma (2015) dalam penelitiannya yang berjudul "Penerapan *K-Means Clustering* Pada Data Penerimaan Mahasiswa Baru (Studi Kasus : Universitas Potensi Utama)", penulis

dalam penelitian ini menjelaskan bahwa hasil dari *cluster* dipengaruhi dari nilai *centroid* awal yang dipakai dan jumlah data yang dipakai, perbedaan pengambilan data pusat *centroid* awal yang dipakai juga akan mempengaruhi hasil *centroid* akhirnya.

II.2. Uraian Teoritis

II.2.1. *Knowledge Discovery in Database*

Menurut Aulia Fitrul Hadi (2017) dalam jurnalnya ia menyatakan bahwa KDD adalah pencabutan yang trivial informasi implisit, yang sebelumnya tidak diketahui, dan berpotensi berguna dari data. KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data. KDD adalah analisis eksplorasi secara otomatis dan pemodelan *repository* data yang besar**.**

Adapun tahap-tahap dari *data mining* ada 7 yaitu :

1. Pembersihan data (*data cleaning*), Pembersihan data merupakan proses penghilangan data yang tidak sesuai atau data yang tidak relevan. Pada umumnya data yang diperoleh, baik dari *database* suatu perusahaan maupun hasil eksperimen. Maka dari itu diperlukannya pembersihan data yang tidak diperlukan.
2. Integrasi data (*data integration*), Integrasi data merupakan gabungan beberapa *database* ke dalam sebuah *database* yang baru. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya.

3. Seleksi Data (*Data Selection*), Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.
4. Transformasi data (*Data Transformation*), Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan.
5. Proses *mining*, Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.
6. Evaluasi pola (*pattern evaluation*), Untuk mengidentifikasi pola-pola menarik kedalam *knowledge based* yang ditemukan. Dalam tahap ini hasil dari teknik *data mining* berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai.
7. Presentasi pengetahuan (*knowledge presentation*), Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses *data mining* adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat

II.2.2. *Data Mining*

Menurut Yuli Asriningtias, Rodhyah Mardhiyah, 2014 dalam jurnalnya yang berjudul “Aplikasi *Data Mining* Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa” menyatakan bahwa *Data Mining* merupakan kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya. Menurut

Alfannisa Annurullah Fajrin dan Algifanri Maulana, 2018 dalam jurnalnya yang berjudul "Penerapan Data Mining Untuk Analisis Pola Pembelian Konsumen Dengan Algoritma *FP-Growth* Pada Data Transaksi Penjualan *Spare Part* Motor" menyatakan bahwa *Data Mining* bukanlah suatu bidang yang baru, tujuan *data mining* adalah bertujuan untuk memperbaiki teknik tradisional sehingga bisa menangani jumlah data yang sangat besar, dimensi data yang tinggi serta data yang heterogen dan berbeda sifat.

Secara sederhana dapat diartikan bahwa *Data Mining* adalah suatu proses penambangan data yang sudah dikumpulkan sebelumnya dalam jumlah yang besar sehingga diperoleh beberapa informasi yang dapat berguna bagi masyarakat ataupun organisasi.

Menurut Pramudiono, 2007, *Data Mining* sebenarnya mempunyai akar yang panjang dari berbagai bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan *database*. Beberapa metode yang paling sering disebut-sebut dalam literatur *data mining* ialah *clustering, classification, association rules mining, neural network, genetic algorithm* dan lain-lain.

II.2.3. *Clustering*

Menurut Ade Bastian, dkk (2018) pada jurnal nya yang berjudul "Penerapan Algoritma K-Means *Clustering* Analisis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka)" menyatakan bahwa pada dasarnya *clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (*similarity*) antar satu data dengan data yang lain.

Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan (*unsupervised*), maksudnya metode ini diterapkan tanpa adanya latihan (*training*) dan tanpa ada guru serta tidak memerlukan target output.

II.2.4. Algoritma K-Means

Menurut Yudi Agusta, PhD dalam jurnalnya yang berjudul "*K-Means – Penerapan, Permasalahan dan Metode Terkait"* menyatakan bahwa *K-Means* merupakan salah satu metode data *clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/kelompok. Metode ini mempartisi data ke dalam *cluster*/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain.

Langkah-langkah yang dilakukan oleh algoritma metode K-Means adalah sebagai berikut : (Budi Santosa, dkk, 2007)

1. Menentukan banyaknya *cluster* yaitu dari data-data yang ada.
2. Pengesetan nilai awal titik tengah/centoid.
3. Menentukan pusat *cluster* secara acak pada data awal.
4. Menghitung data penyakit ke centroid dengan menggunakan rumus jarak Euclid.

   $$d(P,Q) = \sqrt{\sum_{j=1}^{p}(x_j(P) - x_j(Q))^2} \text{.....[1]}$$

5. Melakukan *clustering* data dengan memasukkan setiap objek ke dalam *cluster* (grup) berdasarkan jarak minimumnya.

6. Jika ada data yang harus dipindah, maka langkah selanjutnya adalah menghitung pusat *cluster* baru. Pusat *cluster* yang baru ditentukan berdasarkan pengelompokkan anggota masing-masing *cluster* baru. Pusat *cluster* baru untuk *cluster* yang pertama dihitung berdasarkan ratarata koordinat. Pengulangan dihentikan sampai hasil perhitungan menunjukkan adanya angka pusat *cluster* yang sama.

II.2.5. *Gadget*

Menurut Wikipedia, *Gadget* adalah suatu peranti atau instrumen yang memiliki tujuan dan fungsi praktis yang secara spesifik dirancang lebih canggih dibandingkan dengan teknologi yang diciptakan sebelumnya. Menurut Kurniawan (Rohman 2017: 27) yang dimaksud dengan *gadget* (*smartphone*) yaitu: *Gadget* adalah sebuah perangkat atau perkakas mekanis yang mini atau sebuah alat yang menarik karena relatif baru sehingga akan banyak memberikan kesenangan baru bagi penggunanya walaupun mungkin tidak praktis dalam penggunaannya.

Menurut Fahdian Rahmandani, Agus Tinus, M. Mansur Ibrahim, 2018 dalam jurnalnya yang berjudul "Analisis Dampak Penggunaan *Gadget (Smartphone)* Terhadap Kepribadian dan Karakter (Kekar) Peserta Didik Di SMA Negeri 9 Malang" menyatakan bahwa, *Gadget (smatrphone)* adalah sebuah teknologi yang banyak disukai oleh pemuda bahkan seluruh kalangan di Indonesia maupun dunia yang semakin mempermudah kegiatan informasi dan komunikasi manusia.

II.2.6. *Microsoft Excel*

Microsoft Excel adalah salah satu program aplikasi yang disebut *spreadsheet*, yang memungkinkan pengguna untuk menyediakan data dan instruksi dalam bentuk perintah dan formula untuk membuat persengketaan yang diinginkan. *Spreadsheet* bukan bahasa komputer yang digunakan untuk menulis sebuah program; ini adalah program aplikasi yang dengannya pengguna dapat mengatur prosedur untuk membuat perhitungan dalam bentuk tabel. (Yosaphat Sumardi, 2002)**.**

II.2.7. *RapidMiner*

Menurut Dennis Aprilla C, dkk dalam bukunya yang berjudul "Belajar data mining dengan *rapidminer*" menjelaskan bahwa *RapidMiner* merupakan sebuah perangkat lunak yang bersifat terbuka (*open source*). *RapidMiner* merupakan sebuah solusi dalam melakukan analisis terhadap *data mining*, teks mining dan analisis prediksi. *RapidMiner* merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri.

Dibawah ini merupakan beberapa fitur dari *RapidMiner*, antara lain sebagai berikut:

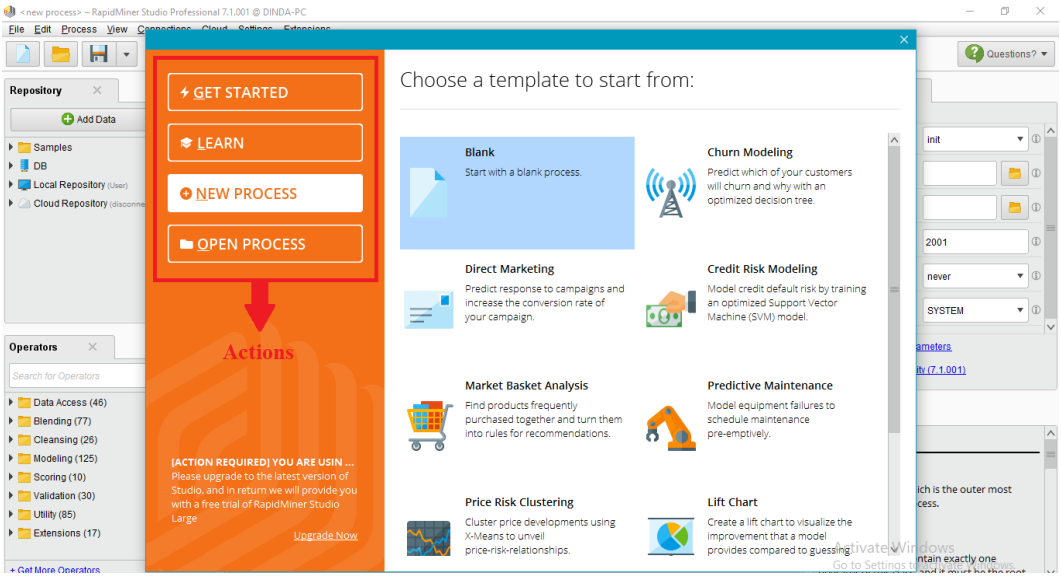
1. Banyaknya algoritma *data mining*, seperti *decision tree* dan *sel-organization map.*
2. Bentuk grafis yang canggih, seperti diagram histogram, *tree chart* dan 3D *scatter plots*.
3. Banyaknya variasi *plugin*, seperti *text plugin* untuk melakukan analisis teks.

4. Menyediakan prosedur *data mining* dan *machine learning* termasuk ETL (*Extraction, Transformation, Loading*), data *preprocessing*, visualiasi, *modelling* dan evaluasi.
5. Proses *data mining* tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI.
6. Mengintegrasikan proyek data mining Weka dan Statistika R.

II.2.7.1 Pengenalan *RapidMiner*

1. Tampilan Awal RapidMiner

   Tampilan utama *tools RapidMiner* dapat dilihat pada gambar dibawah ini.



Gambar II.1. Tampilan Awal *RapidMiner*

Bagian **Actions** menunjukkan daftar aksi yang dapat di lakukan setelah membuka *RapidMiner*. Berikut ini rincian lengkap daftar aksi tersebut:

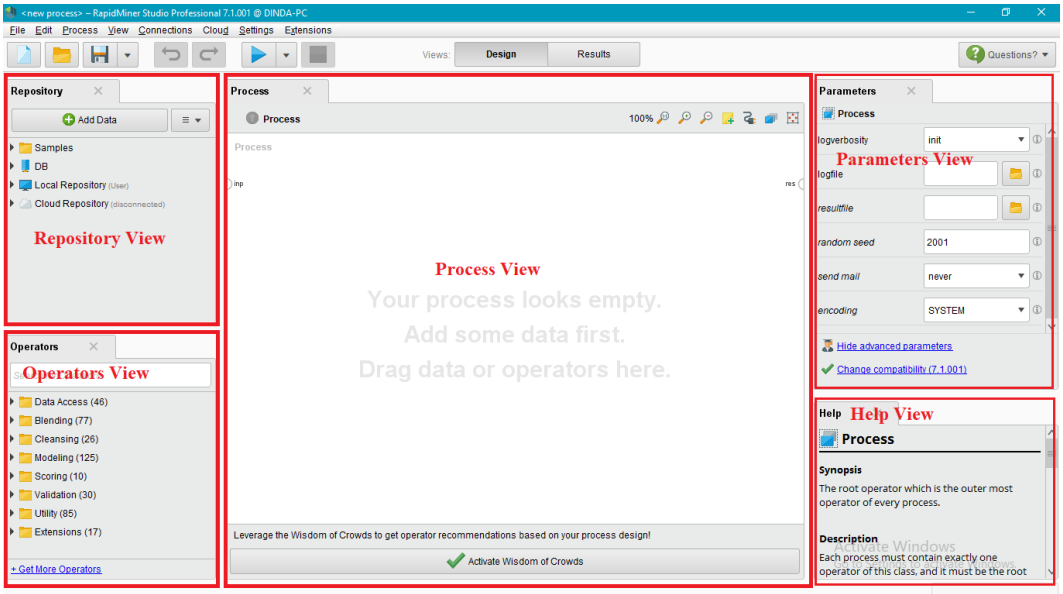
a. ***New Process*** : Aksi ini berguna ntuk memulai proses analis baru. Untuk memulai proses analisis, pertama-tama harus menentukan nama dan lokasi proses dan *Data Repository*. Setelah itu, bisa mulai merancang sebuah analisis baru.

b. ***Learn :*** Aksi digunakan untuk memulai tutorial secara *online* (terhubung internet). Tutorial yang dapat secara langsung digunakan dengan *RapidMiner* ini, memberikan perkanalan dan beberapa konsep data mining. Hal ini direkomendasikan untuk yang sudah memiliki pengetahuan dasar mengenai data mining dan sudah akrab dengan operasi dasar *RapidMiner*.

c. ***Open Process*** : Aksi ini untuk membuka *Repository Browser* yang berisi daftar proses. Anda juga bisa memilih proses untuk dibuka pada *Design Perspective.*

2. Tampilan Utama *RapidMiner*

Tampilan Utama *RapidMiner* dapat dilihat pada gambar dibawah ini.



Gambar II.2. Tampilan Utama *RapidMiner*

Sebagai lingkungan lingkungan kerja, Design Perspective memiliki beberapa view. Berikut ini beberapa view yang ditampilkan pada Design Perspective:

1. *Operator View*

Operator View merupakan view yang paling penting pada perspective ini. Semua operator atau langkah kerja dari *RapidMiner* disajikan dalam bentuk kelompok hierarki di *Operator View* ini sehingga operator-operator tersebut dapat digunakan pada proses analisis. Hal ini akan memudahkan dalam mencari dan menggunakan operator yang sesuai dengan kebutuhan. Pada *Operator View* ini terdapat beberapa kelompok operator sebagai berikut:

- *Process Control* : Operator ini terdiri dari operator perulangan dan percabangan yang dapat mengatur aliran proses.
- *Utility* : Operator bantuan, seperti operator macros, login, subproses, dan lain-lain.
- *Repository Access* : Kelompok ini terdiri dari operator-operator yang dapat digunakan untuk membaca atau menulis akses pada *repository*.
- *Import* : Kelompok ini terdiri dari banyak operator yang dapat digunakan untuk membaca data dan objek dari format tertentu seperti file, database, dan lain-lain.
- *Export* : Kelompok ini terdiri dari banyak operator yang dapat digunakan untuk menulis data dan objek menjadi format tertentu.
- *Data Transformation* : kelompok ini terdiri dari semua operator yang berguna untuk transformasi data dan meta data.

- *Modeling* : kolompok ini berisi proses data mining untuk menerapkan model yang dihasilkan menjadi set data yang baru.
- *Evaluation* : kelompok ini berisi operator yang dapat digunakan untuk menghitung kualitas pemodelan dan untuk data baru.

2. *Repository View*

Repository View merupakan komponen utama dalam *Design Perspective* selain *Operator View*. View ini dapat gunakan untuk mengelola dan menata proses Analisis menjadi proyek dan pada saat yang sama juga dapat digunakan sebagai sumber data dan yang berkaitan dengan meta data.

3. *Process View*

Process View menunjukkan langkah-langkah tertentu dalam proses analisis dan sebagai penghubung langkah-langkah tersebut. dapat menambahkan langkah baru dengan beberapa cara. hubungan diantara langkah-langkah ini dapat dibuat dan dilepas kembali. Pada dasarnya bekerja dengan *RapidMiner* ialah mendefinisikan proses analisis, yaitu dengan menunjukkan serangkaian langkah kerja tertentu. Dalam *RapidMiner*, komponen proses ini dinamakan sebagai operator. Operator pada *RapidMiner* didefinisikan sebagai berikut:

- Deskripsi dari input yang diharapkan.
- Deskripsi dari output yang disediakan.
- Tindakan yang dilakukan oleh operator pada input, yang akhirnya mengarah dengan penyediaan output.
- Sejumlah parameter yang dapat mengontrol *action performed*.

4. *Parameter View*

Beberapa operator dalam *RapidMiner* membutuhkan satu atau lebih parameter agar dapat diindikasikan sebagai fungsionalitas yang benar. Namun terkadang parameter tidak mutlak dibutuhkan, meskipun eksekusi operator dapat dikendalikan dengan menunjukkan nilai parameter tertentu. *Parameter view* memiliki *toolbar* sendiri sama seperti view-view yang lain. pada Parameter View ini terdapat beberapa ikon dan nama-nama operator terkini yang dikuti dengan aktual parameter.

Huruf tebal berarti bahwa parameter mutlak harus didefinisikan oleh analis dan tidak memiliki nilai default. Sedangkan huruf miring berarti bahwa parameter diklasifikasikan sebagai parameter ahli dan seharusnya tidak harus diubah oleh pemula untuk analisis data. Poin pentingnya ialah beberapa parameter hanya ditunjukkan ketika parameter lain memiliki nilai tertentu.

5. ***Help View***

Setiap kali memilih operator pada *Operator View* atau *Process View*, maka jendela bantuan dalam *Help View* akan menunjukkan penjelasan mengenai operator ini. Penjelasn yang ditampilkan dalam *Help View* meliputi:

- Sebuah penjelasan singkat mengenai fungsi operator dalam satu atau beberapa kalimat.
- Sebuah penjelasan rinci mengenai fungsi operator.
- Daftar semua parameter termasuk deskripsi singkat dari parameter, nilai default (jika tersedia), petunjuk apakah parameter ini adalah parameter ahli serta indikasi parameter dependensi.