BAB II

LANDASAN TEORI

2.1. Tinjauan Pustaka

Penelitian dan penerapan terkait *data mining* juga sudah banyak dilakukan oleh perusahan-perusahaan, intansi pemerintahan dan pendidikan. Referensi dan rujukan terhadap hasil penelitian sebelumnya yang berhubungan dengan penelitian yang dilakukan adalah tujuan dari tinjauan pustaka ini.

Menurut Indraloka Smaradahana, Santosa (2017) melakukan penelitian dengan judul "Penerapan *Text Mining* Untuk Melakukan *Clustering* Data *Tweet* Shopee Indonesia". Dalam penelitiannya pengolahan dan pengambilan informasi dari sebuah data tekstual dan dilakukan secara *real time* dari Twitter, karena permasalahan ini tidak dapat diselesaikan dengan menggunakan teknik *Data Mining.* Penerapan *Text Mining* untuk melakukan *clustering* dengan metode *K-means*, penentuan jumlah *cluster* dilakukan berdasarkan perhitungan nilai *Silhouette Coefficient* berdasarkah hasil perhitungan diketahui ada sebanyak 21 *cluster* yang memiliki nilai positif, 3 *cluster* memiliki nilai 0, dan 4 *cluster* memiliki nilai negatif. Tidak dipungkiri masih ada beberapa faktor yang membuat hasil *clustering* dengan metode *k-means* masih belum optimal. Konten *tweet* yang banyak dilakukan *retweet* oleh *followers* dapat mempengaruhi para pelaku bisnis Shopee Indonesia sebagai sarana untuk *advertising* kepada pengguna Twitter.

Menurut penelitian Asroni, Adrian (2015) telah melakukan penelitian dengan judul "Penerapan Metode *K-Means* Untuk *Clustering* Mahasiswa Berdasarkan Nilai Akademik Dengan *WEKA Interface* Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang". Penelitian ini mencari 5 orang mahasiswa jurusan Teknik Informatika untuk mengikuti sebuah lomba, adapun lomba yang akan diikuti adalah kompetisi *even Cyberjawara* yang diselenggarakan oleh *indonesia security incident response team on internet infrastructure* (ID SIRTII) Kementrian Komunikasi dan Informatika RI.

Penelitian dilakukan menggunakan data yang telah ada di *data warehouse* UMM Magelang. Pengujian ini memiliki 5 atribut yaitu nim mahasiswa, nilai matakuliah algoritma dan pemograman 1, nilai matakuliah fisika dasar, nilai kalkulus 1, IPK dan jumlah *instance* adalah 124. Menggunakan sofware *WEKA,* tujuannya untuk membandingkan hasil perhitungan teoritis dengan hasil yang menggunakan proses perhitungan di *WEKA Interface.* Menggunakan persamaan *Euclidean. Cluster* 0 dengan IPK = 0,5167 sebanyak 9 mahasiswa (7%), *cluster* 1 dengan IPK = 3,4143 sebanyak 28 mahasiswa (23%), *cluster* 2 dengan IPK = 3,3092 sebanyak 40 mahasiswa (32%) dan *cluster* 3 dengan IPK = 3,8991 sebanyak 47 mahasiswa (38%). Maka *cluster* 1 dengan IPK tertinggi. Dari data tersebut didapat 4 kelompok, dan *cluster* 1 bisa digunakan untuk memilih 5 mahasiswa untuk mewakili mengikuti lomba.

Penelitian yang dilakukan oleh Mulida (2018) dengan penelitiannya berjudul "Penerapan Datamining Dalam Mengelompokkan Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Prov. DKI Jakarta Dengan *K-Means*". Ada 8 objek wisata DKI Jakarta dan tujuan dari penelitian ini adalah menganalisis penerapan *datamining* dan mengelompokkan jumlah wisatawan asing ke Prov. DKI Jakarta. Dalam hal ini ada 3 jumlah *cluster,* yaitu (C1) untuk *cluster* tinggi, (C2) untuk *cluster* sedang dan (C3) untuk *cluster* rendah. Sehingga diperoleh hasil dari *K-Means* bahwa C1 terdiri dari 1 objek wisata unggulan Taman Impian Jaya Ancol, C2 terdiri dari 2 objek wisata yakni Taman Mini Indonesia dan Kebon Binatang Ragunan, sisanya adalh 5 objek wisata yang masuk pada C3. Perbaikan sarana dan prasana akan meningkatkan jumlah kunjungan wisatawan.

Carvalho, Goncalves (2016) melakukan penelitian yang berdasarkan pada perubahan iklim dari berbagai variabel. yang berjudul "*Regionalization of Europe of Europe based on a K-Means Cluster Analysis of the climate change of temperatures and precipitation*". Dengan menerapkan analisis pengelompokan menggukan *k-means* dengan perbedaan iklim harian untuk setiap variabel, diutamakan pada perubahan jangka panjang dalam curah hujan, suhu maksimum dan suhu minimum. Hasilnya dari menggunakan setiap variabel sebagai fitur (versi

multivariate) adalah peta dimana setiap titik *grid* untuk *cluster* (wilayah). Perbedaan klimatologi musiman rata-rata untuk masing-masing daerah menjadi jelas ketika dianalisis, bahwa daerah tersebut memiliki musiman rata-rata. Pada kenyataannya, curah hujan, suhu maksimum dan suhu minimum mempunyai karakteristik yang berbeda ketika diproyeksikan perubahan. Hasil penelitian menunjukkan bahwa, secara statistik setiap variabel yang yang pasangan wilayah dibandingkan tidak signifikan, ketika meningkatkan jumlah *cluster* dianggap ada peningkatan rinci dalam fitur yang diperoleh. Iklim musiman ini terdeteksi perubahannya tidak jelas pada fungsi probalitas distribusi variabel asli dan beberapa daerah ditemukan tidak signifikan. Perbedaan satu sama lain mengenai perubahan variabel, cara untuk mendekati subjek mengidentifikasi daerah perubahan iklim koheren metodologi ini menyajikannya menggunakan sebuah novel. Selain itu, bukan hanya satu variabel yang digunakan untuk menentukan dan menciptakan daerah .

Pengujian yang dilakukan iterasi *clustering* data mahasiswa adalah sebanyak 2 kali. Kesimpulan yang dapat diambil adalah jika asal Sekolah dalam Sekolah Menengah Pertama maka rata-rata jurusan yang diambil adalah Sistem Informasi dan jika alas Sekolahnya adalah SMK rata-rata yang dipilih adalah Teknik Informatika. *Cendroid* awal juga mempengaruhi hasil *cluster* dan jumlah data yang dipakai, hasil *cendroid* akhir juga di pengaruhi oleh pengambilan data yang berbeda pada *cendroid* awal. Nasari, Darma (2015) penelitian mereka yang berjudul "Penerapan *K-Means Clustering* Pada Data Penerimaan Mahasiswa Baru (Studi Kasus: Universitas Potensi Utama)".

Penelitian yang berjudul "Implementasi Algoritma *K-Means* Untuk Pengelompokan Penyakit Pasien Pada Puskesmas Kajen Pekalongan" yang dilakukan oleh Wardhani, Khrisna (2016). Data yang digunakan pada penelitian ini diperoleh dari data pasien Puskesmas Kajen Pekalongan, kemudian diolah menggunakan *K-Means* untuk mengelompokkan data penyakit pasien dibagi menjadi 2 *cluster* yaitu (C1) untuk "akut" berjumlah 376 *item*, dan (C2) untuk "tidak akut" berjumlah 624 *item* dengan total jumlah data 1000. Karena

pengelolahan data kesehatan ini masih menggunakan perhitungan manual sehingga menghasilkan *output* yang kurang maksimal dan memiliki permasalahan pada konsistensi pada data, oleh sebab itu untuk menentukan konsistensi data kesehatan dapat digunakan teknik *datamining* yang mampu menggali informasi yang tersembunyi dari kumpulan data.

Menurut peneliti Aranda, Natasya (2016) "Penerapan Metode *K-Means Cluster Analysis* Pada Sistem Pendukung Keputusan Pemilihan Konsentrasi Untuk Mahasiswa *International Class* STMIK AMIKOM Yogyakarta". Penelitiannya menguji iterasi *clustering* data mahasiswa sebanyak 3 kali iterasi maka ditemukan hasil 4 dari 12 mahasiswa diarahkan untuk mengambil konsentrasi Pemograman dan 4 mahasiswa mengambil konsentrasi Multimedia dan 3 – 5 mahasiswa mengambil konsentrasi jaringan Komputer. Nilai *cendroid* awal yang dipakai dan jumlah data sangat mempengaruhi hasil *clusternya,* selain itu perbedaan antara pengambilan data pusat *cendroid* awal juga mempengaruhi hasil pada *cendroid* akhir.

Penelitian yang dilakukan oleh Ong (2013) dengan judul "Implementasi Algoritma *K-Means Clustering* Untuk Menentukan Strategi Marketing President University". melakukan promosi yang dilakukan oleh pihak *marketing President University* dalam penelitian ini dari hasil *clustering* agar lebih efektif dan effisien. Berdasarkan hasil pengelompokan data menggunakan *K-means* didapat 3 *cluster.*1 rata-rata mahasiswa mengambil jurusan Information Technology dan Marketing yang berasal dari daerah DKI Jakarta dan Jawa Barat. *Cluster* 2 berasal dari kota DKI Jakarta dan Jawa Barat mengambila jurusan Information Technology dan Marketing. Dan yang terakhir adlaah *cluster 3* berasal dari daerah Sulawesi, Jawa Timur dan Sumatra Selatan dengan mengambil jurusan Public Relation, Accounting dan International Business. Kesimpulan dari penelitian ini adalah pihak *marketing President University* melakukan promosi mengirim tim yang sesuai dengan jurusan yang paling banyak diminati pada kota-kota berdasarkan tingkat memapuan akademik dari calon mahasiswa.

Penelitian yang berjudul "*Clustering Analysis on Alumni Data Using Abandoned and Reborn Particle Swarm Optimization* " dari Mudjihartono, Tanprasert, Jiamthapthaksin (2016). Dalam PSO (*Particle Swarm Optimization*) *centroid* memecahkan beberapa partikel, partikel ini bergerak dalam iterasi sebagai PSO dasar. Fungsinya adalah untuk meminimalisir kesalahan *cluster* . menggunakan 3 pendekatan dan pedekatan terakhir untuk menghubungkan *cluster* dengan fakta-fakta lebih bermakna tentang data. Secara khusus, itu membandingkan kemurnian pengelompokan mengenai kebenaran dasar data. SEE meringkas semua kesalahan kuadrat dari semua *cluster*.

Penelitian yang dilakukan oleh Ji, Pang, Zheng, Wang, Ma (2015) dalam penelitiannya ini mengusulkan sebuah novel koloni lebah pendekatan *clustering* buatan untuk data kategorikal. ABC-K-Modes (*Artifical Bee Colony* pengelompokkan berdasarkan *K-Mode*), dalam algoritma ini kawanan lebah buatan terdiri dari 3 jenis yaitu lebah pekerja, lebah penonton, dan lebah pramuka. Tiap lebah memiliki tugas masing-masing, lebah penonton bertugas untuk mengambil sumber makanan tertentu untuk mengeksploitasi dan bagai informasi tentang sumber makanan dengan menonton dari dalam sarang, lebah pramuka mencari sumber makanan dalam ruang pencarian, lebah pekerja bekerja sama dengan lebah penonton untuk mendapatkan makanan dari informasi yang diberikan oleh lebah pramuka. Kompleksitas metode yang diusulkan terditi dari 5 bagian, yaitu inisialisasi, operasi pencarian lebah yang bekerja, perhitungan probabilitas sumber makanan, dan operasi pencarian pengintai dan penonton. Dalam algoritma ini, proses pencarian lebah pekerja dan lebah penonton diimplementasikan dengan memperkenalkan prosedur khusus bernama *OKM,* dan proses pencarian lebah pramuka dilakukan secara acak.

2.2. Landasan Teori

2.2.1. Data Mining

Pattern recognition atau *data mining* merupakan metode yang digunakan untuk pengolahan data guna menentukan pola yang tersembunyi dari data yang diolah. Hasil data yang diolah menggunakan *data mining* ini berupa pengetahan baru yang bersumber dari data lama, dapat digunakan dalam menentukan keputusan di masa depan (Ong, Johan Oscar 2013).

Data mining memiliki peranan yang sangat penting dalam beberapa bidang kehidupan karena merupakan metode yang digunakan dalam pengolahan data berskala besar. *Data mining* juga memiliki beberapa metode antara lain klasifikasi, *clustering, regresi,* seleksi variabel, dan *market basket analisis* (Ong, Johan Oscar 2013).

Data mining dalam aplikasinya merupakan salah satu bagian proses *Knowledge Discovery in Database* (KDD) yang memiliki tugas mengekstrak sebuah pola atau model dari data dengan menggunakan suatu algoritma yang spesifik. Proses KDD sebagai berikut: (Angga Ginanjar Mabrur 2012).

1. *Data selection.*

   Pemilihan data / seleksi data dari sekumpulan data operasional perlu dilakukan sebelum masuk ke tahap penggalian informasi dalam KDD. Hasil data yang sudah diseleksi akan digunakan untuk proses *data mining,* dan disimpan dalam suatu berkas yang terpisah dari basis data operasisonal.

2. *Pre-processing.*

   Ada tahapan sebelum masuk pada proses *data mining* yaitu proses *pre-processing* atau yang sering disebut *cleaning.* Yang bertujuan untuk membuang duplikasi data, memastikan data yang *inkosisten,* dan memperbaiki kesalahan data, seperti kesalahan cetak / *tipografi. Enrichement* adalah proses "memperkaya" data yang sudah ada dengan informasi lain yang relevan dan di perlukan untuk KDD.

3. *Transformation.*

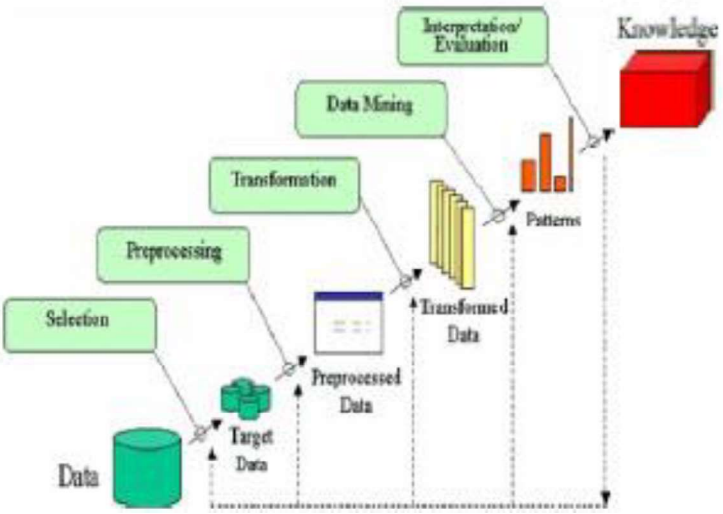
   Proses *coding* pada data yang telah dipilih adalah transformasi. Data diproses hingga sesuai untuk *data mining.* Dalam KDD ada proses *coding* yang kreatif dan sangat tergantung pada jenis pola informasi yang akan dicari dalam basis data.

4. *Data mining.*

   Data mining adalah sebuah proses mencari pola atau informasi menarik dalam data yang terpilih dengan menggunakan metode tertentu. Metode atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat tergantung tujuan dan proses KDD secara keseluruhan.

5. *Interpretation / Evaluation.*

   Proses *data mining* yang dihasilkan oleh pola informasi perlu ditampilkan dalam bentuk yang mudah dimengerti. Tahap ini disebut *interpretation* yang merupakan bagian dari proses KDD. Pada tahap pemeriksaan ini mencakup pola atau informasi yang ditemukan apakah bertentangan dengan fakta atau hipotesis yang ada sebelumnya (Fina Nasari dan Surya Darma 2015).



Gambar 2. 1 Tahapan data mining

Menurut (Linda Maulida 2018) *data mining* adalah suatu istilah yang digunakan untuk menukan sebuah pengetahuan dalam *database* seperti pada gambar 2.1. Proses *data mining* menggunakan teknik statik, metamatika, kecerdasan buatan, dan *machine learning* untuk mengeluarkan dan menganalisa informasi yang bermanfaat terkait pengetahuan dari berbagai *data base* besar.

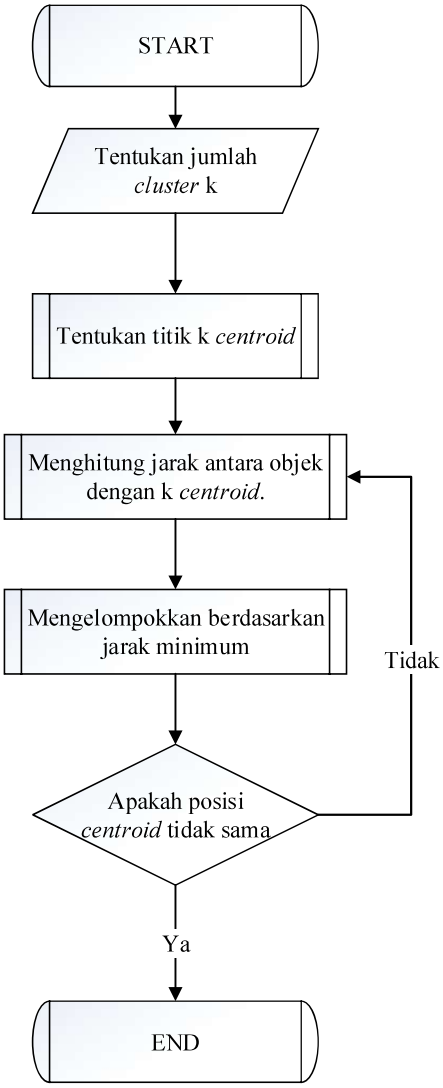
2.2.2. *Clustering*

Clustering adalah seuatu metode untuk mencari dan mengelompokkan data yang memiliki persamaan karakteristik / *similarity* antara satu data dengan data yang lainnya. (Ong, Johan Oscar 2013) *Clustering* merupakan metode yang bersifat tanpa arah / *unsupervised,* maksudnya adalah metode ini diterapkan tanpa adanya latihan / *taining* dan tanpa ada guru / *teacher* serta tidak membutuhkan target *output.*

Pengelompokan data-data kedalam beberapa jumlah kelompok / *cluster* berdasarkan struktur kelompok, keanggotaan data dalam kelompok, dan kekompakan data dalam kelompok. Ada dua jenis metode menurut pengelompokan, yaitu hierarki dan *partitioning.* Untuk hierarki, Satu data tunggal bisa dianggap sebuah kelompok, dua atau lebih. Sedangkan untuk *partitioning* membagi data menjadi satu kelompok. Jika menurut keanggotaan data, dibagi menjadi dua, yaitu ekslusif dan tumpang-tindih. Kategori ekslusif, data dipastikan hanya dapat menjadi satu kelompok. Kategori tumpang-tindih data dapatmenjadi anggota lebih dari satu kelompok. Berdasarkan kategori kekompakan ada komplet dan parsial. Data yang menyimpang disebut *outlier, noise,* atau *uninterested background.*

2.2.3. *K-Means*

K-means adalah metode pengelompokkan data *non-hierarki* yang mempartisi data dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data yang berkarakter sama dan dikelompokkan menjadi satu dan kemudian data yang berkarakter beda akan dikelompokkan ke dalam kelompok lain. Tujuan pengelompokkan ini adalah untuk meminimalkan fungsi objektif yang ada di *set* dalam suatu kelompok dan memaksimalkan variasi antar kelompok.



Gambar 2. 2 Diagam Alir K-Means

Pengertian dari *k-means Clustering* adalah, *k* sebagai konstanta jumlah *cluster* yang diinginkan, *Means* artinya nilai rata-rata dari suatu grup data yang didefinisikan sebagai *cluster*, sehingga *k-means clustering* adalah metode penganalisa *data mining* yang melakukan proses pemodelan tanpa supervisi / *unsupervised* dan merupakan salah satu metode yang melakukan pengelompokkan data dengan sistem partisi. Metode *k-means* berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, data dalam satu kelompok mempunyai karakter yang sama satu sama lainnya dan data yang memiliki perbedaan karakteristik akan

dipisahkan dalam kelompok yang berbeda. Dasar algoritma *k-means* adalah sebagai berikut:

1. Tentukan nilai *k* sebagai jumlah *cluster* yang akan dibentuk.
2. Inisialisasi *k* sebagai *centroid* yang dapat dibangkitkan secara *random.*
3. Hitung jarak setiap data ke masing-masing *centroid* menggunakan persamaan *Euclidean Distance*

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

Keterangan:

d = data titik dokumen (*euclidean*)

P = data *record*

Q = data *centroid*

4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroidnya*.
5. Tentukan posisi *centroid* baru *(k)*.
6. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.

2.2.4. RapidMiner Studio

RapidMiner Studio merupakan perangkat lunak yang bersifat terbuka (*open source*) yang memberikan solusi untuk melakukan analisis pada *data mining*, *text mining* dan analisis prediksi. *RapidMiner Studio* menggunakan berbagai macam cara deskriptif dan prediksi dalam memberikan pengetahuan pada pengguna sehingga dapat membuat keputusan yang terbaik. *RapidMiner Studio* memiliki kurang lebih 500 operator *data mining*, termasuk operator *input, output, data preprocessing* dan visualisasi. *RapidMiner Studio* ini diprogram menggunakan bahasa *java* sehingga dapat bekerja pada semua sistem operasi.

2.2.5. Microsoft Excel

Microsoft excel adalah *software spreadsheet* paling terkenal didunia bisnis, perkantoran maupun pendidikan. *Excel* sangat membantu banyak pekerjaan disetiap bidang, dan selalu dijumpai dimanapun karena aplikasi ini sangat universal. *Software excel* ini memiliki banyak fitur kalkulasi dan pembuatan grafis, serta mudah digunakan menjadikan *excel* sebagai *software* paling banyak diminati para pengguna terutama pekerja.