BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian terdahulu adalah upaya peneliti untuk menemukan perbandingan untuk penelitian selanjutnya. Di samping itu ini berguna untuk membantu peneliti dapat memposisikan penelitian nuntuk menyajikan hasil penelitian yang orisinil. Dalam situasi ini, peneliti mengumpulkan berbagai hasil studi yang terkait dan kemudian membuat ringkasan dari hasil tersebut. Berikut ini adalah kumpulan esai yang sebagian besar terkait dengan topik yang sedang ditulis secara rinci yang dijabarkan pada tabel 2.1

Tabel 2.1 Penelitian Terdahulu

Penulis	Judul	Hasil Penelitian	Akurasi
Aiman ayadi, kusrini, eko pramono	Perbandingan tingkat performa metode K-Means dan hierachial clustering pada sistem rekomendasi pemilihan kost	Kesimpulan hasil perbandingan tingkat performa pada penelitian ini adalah model algoritma K-Means + Naive Bayes memiliki nilai rata-rata akurasi yang tertinggi, kemudian nilai rata-rata presisi model algoritma algoritma K-Means + Setelah melalui tahap pengujian tingkat peforma dengan parameter akurasi, presisi, recall dan Waktu Training pada model algoritma Hierachical + Naive Bayes serta Hierachical Naive Bayes dan Optimalisasi untuk prediksi mahasiswa pada rekomendasi pemilihan kost.	90,82%
Chandra purmaningsih, ristu saptono, abdul aziz	Pemanfaatan metode K-Means clustering dalam penentuan penjurusan siswa SMA	Dari hasil penelitian, dapat disimpulkan bahwa algoritma K-Means clustering dapat digunakan untuk mengelompokkan data siswa sebagai pendukung keputusan penentuan penjurusan siswa SMA. Berdasarkan hasil pengujian terbaik pada praprocessing clustering K-Means IPA dengan hasil akurasi 0.905882.	90.59%

Penulis	Judul	Hasil Penelitian	Akurasi
Ayu sari nurlatifah, martaleli bettiza, ferdi chahyadi	Implementasi Algoritma clustering DBSCAN untuk menentukan status gizi balita	Penerapan Algoritma Clusteirng DBSCAN untuk menentukan status gizi balita memiliki akurasi tertiggi 52% dan nilai akurasi terendah 11%. Sehingga Algoritma Clustering DBSCAN dapat digunakan untuk menentukan status gizi balita	52%
Syham lal, dr. Vaishali singh	Techniques to enhance the performance of DBSCAN clustering algorithm in data mining	Makalah ini menyajikan studi komprehensif tentang algoritma DBSCAN dan versi yang disempurnakan dari algoritma DBSCAN dengan dengan implementasi menggunakan matlab. Disimpulkan bahwa cluster berbasis densitas adalah jenis cluster yang ekonomis sepanjang unit cluster yang dicetak pada kepadatan dataset input	92%
Tukiyat, makhsun, yohanes, siti hodijah	Klasterisasi persebaran covid-19 di kota tanggerang selatan dengan metode clustering of aplications with noise	Berdasar pada hasil DBI tersebut di atas daapt ditunjukkan bahwa Algoritma K-Means lebih baik dibanding algoritma DBSCAN karena DBI karena nilai DBI pada algoritma K-Mean lebih kecil dibanding dengan DBI pada DBSCAN. Dengan demikian, maka model clusterisasi dengan algoritma K-Means lebih cocok untuk mengkonstruksikan model clusterisasi pademi covid-19 di Kota Tangerang Selatan	67%

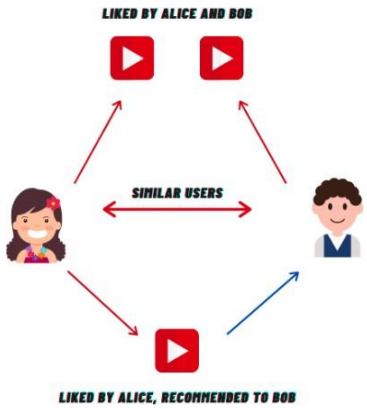
Dalam beberapa penelitian sebelumnya, peneliti menggunakan pendekatan hybrid, CF dan klasifikasi. Pembaharuan yang dilakukan pada penelitian ini adalah penggunakaan algoritma *clustering* dan mengkolaborasikan dengan *Cosine similarity* untuk merekomendasikan produk yang lebih efektif dan sesuai dengan ketertarikan *user*.

2.2 State Of The Art

State of the art memperlihatkan hasil-hasil penelitian yang berkaitan dengan penelitian yang akan diajukan. Bagian ini akan menunjukkan peluang/posisi penelitian Komparasi DBSCAN dan K-Means dalam mengatasi masalah *cold start* Pada *Collaborative Filtering*, sehingga rekomendasi yang dihasilkan lebih akurat atau sesuai dengan peminatan pengguna (*user*).

2.3 Collaborative Filtering

Collaborative filtering merupakan salah satu metode rekomendasi yang menggunakan data rating dari pengguna lain untuk menghasilkan rekomendasi. Metode tersebut menganggap bahwa selera pengguna terhadap suatu produk akan cenderung sama dari waktu ke waktu, begitu pula dengan pengguna lain yang memiliki selera sama. Lebih sederhananya metode *collaborative filtering* bekerja dengan cara memprediksi apa yang akan disukai pengguna, berdasarkan pada kemiripan dengan pengguna lain. Kemiripan tersebut diperoleh dari menganalisis sekumpulan besar informasi tentang perilaku, atau preferensi pengguna (A. Approuch, G et al, 2015).



Gambar 2.1 Ilustrasi Collaborative Filtering

Sebagai contoh pada Gambar 2.1. Teknik CF hanya memerlukan sekumpulan item yang didasarkan oleh *history* pengguna. Sumber tersebut digunakan untuk merekam interaksi antar produk. Bisa melalui video yang di sukai, trek music ataupun pencarian. Dengan melihat *history* yang memiliki kesamaan, maka CF akan menghubungkan pengguna serupa dengan kesamaan dalam preferensi dan perilaku produk serupa saat mengusulkan ke pelanggan lainnya. CF dapat memberikan rekomendasi video selanjutnya yang mungkin akan di sukai oleh Alice maupun Bob. Algoritma tersebut dirancang sedemikian rupa untuk memprediksi minat Bob dan Alice. Pada pendekatan menggunakan metode CF preferensi yang diberikan oleh pengguna dapat bersifat implisat.

2.4 Cold Star

Cold start adalah suatu kondisi pengguna baru yang belum pernah memberikan rating terhadap suatu produk, sehingga informasi yang didapatkan untuk arah peminatan pengguna sulit diketahui. Jika arah peminatan tidak diketahui, maka

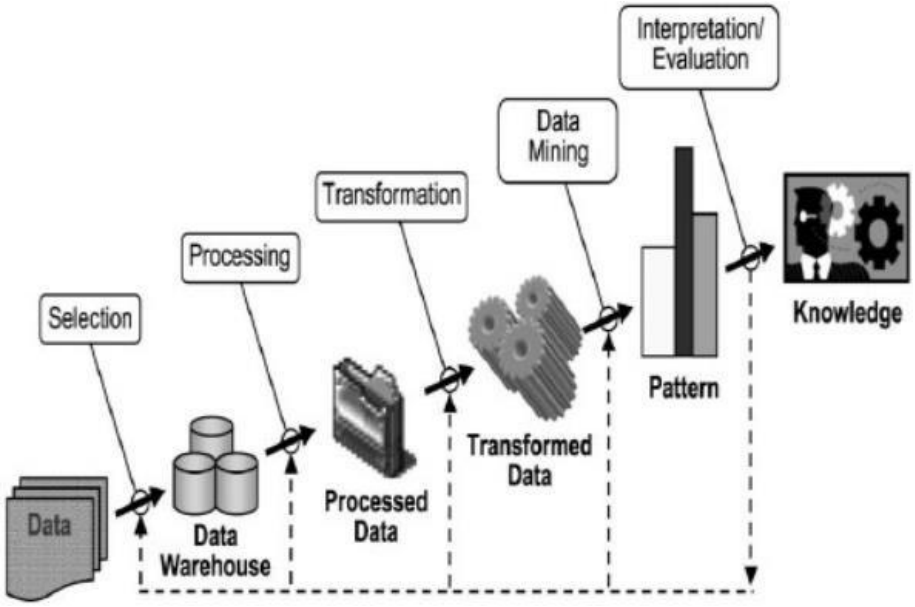
sulit untuk memberikan rekomendasi (S. Sharma & A. Mahajan, 2017). Rating merupakan suatu kegiatan menilai berdasarkan suatu skala satuan dari nilai terendah hingga tertinggi. Pada umumnya, konsumen akan lebih percaya kepada produk/jasa yang sebelumnya telah memiliki banyak ulasan/rating pada produk/jasa sebagai acuan kualitas. Sehingga, produk/jasa yang belum memiliki rating (*cold star*) akan lebih sulit menonjol.

2.5 Data Mining

Data mining adalah sebuah disiplin ilmu yang mempelajari metode untuk mengekstrasi data yang terekam dan tidak memiliki sebuah nilai guna menjadi sebuah pengetahuan. Secara teknis, *data mining* dapat disebut sebagai proses untuk menemukan sebuah pola dari ribuan bahkan jutaan *record* data. Proses pengumpulan dan esktrasi data tersebut dapat menggunakan aplikasi perangkat lunak perhitungan statistika, matematika ataupun *Artificial Intelligent* (AI). Terdapat 5 peran pada data mining, yaitu estimasi, forecasting, asosiasi, klasifikasi dan *clustering*.

Menurut Y Mardi (2017) *data mining* adalah sebuah tahapan dalam *knowledge discovery in database* (KDD). Pada *data mining* kita dapat melakukan pengklasifikasian, prediksi, klasterisasi sehingga dapat menemukan informasi yang bermanfaat dalam kumpulan data yang sangat besar. Ekstrasi pengetahuan yang menarik dari sekumpulan data (D. Agrawal et al, 2011) Sedangkan menurut kusrini (2009) istilah dari *data mining* dan KDD sering kali digunakan secara bergantian untuk menjelaskan lebih lanjut proses penggalian informasi yang tidak nampak

dalam suatu basis data yang besar. Pada saat ini istilah *data mining* memiliki banyak nama. Seperti *big data, business intelegent, pattern analysis, information harvesting, knowledge extraction.* Proses *data mining* memiliki tujuan dimana mengekstrasi data ke pengetahuan yang selanjutnya dapat dilihat pada gambar 2.2 dibawah ini



Gambar 2.2 Konsep proses data mining

Beberapa tahun terakhir, data semakin beragam bentuk dan semakin kompleks dengan jumlah volume yang semakin meningkat cepat. Tidak terbayang seberapa cepat penyebaran data dan data yang diciptakan setiap detiknya oleh manusia. Oleh karena itu munculnya istilah big data yang menggambarkan volume data yang sangat besar yang terstruktur maupun tidak terstruktur. Dalam big data, kita akan mengalami kesulitan dalam membaca dan mengetahui pola-pola dan relasi data jika dilakukan secara manual dan konvensional. Sebagai contoh, pelanggan sebuah layanan video streaming dengan ratusan juta pelanggam yang tersebar di seluruh

Indonesia. Dalam setahun, pelanggan tersebut dapat menghasilkan milyaran data dengan memberikan rating terhadap sebuah film, menonton beberapa buah film dengan berbagai genre, film apa yang di simpan kedalam playlist, film yang tidak disukai dan masih banyak lagi. Lalu muncul pertanyaan bagaimana menentukan pola ketertarikan konsumen terhadap film yang di rekomendasikan? Dengan menggunakan *data mining* pola ketertarikan konsumen dapat diklasifikasikan sedemikian rupa.

2.5.1 Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dala *knowledge data discovery* (KDD) dimulai. Data hasil seleksi yang kelak digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data operasional. Pada tahap ini kita dapat memilih himpunan dari sebuah data atau dapat juga untuk memfokuskan pada subset variable data yang nantinya akan di proses lebih lanjut. Proses tersebut harus relevan terhadap analisis agar data tersebut selaras dengan dataset yang ada.

2.5.2 *Preprocessing* atau Cleaning

Sebelum proses *data mining* dapat dilaksanakan, diperlukan proses *preprocessing* data. Proses ini mencakup membuang atau menghapus data duplikat, memeriksa data yang tidak konsisten dan melakukan pengecekan data. Proses penghapusan data tergantung kepada cara penyimpanannya dan bentuk jawaban yang dicari. *Data cleansing* tidak hanya seputar menghapus data ataupun memberi ruang kepada

data baru, pada tahap ini juga kita dapat menemukan sebuah cara baru untuk memaksimalkan hasil dari akurasi tanpa menghapus informasi apapun. Memperbaiki kesalahan penulisan sintaks, kesalahan ejaan, menstandarisasi kumpulan dari sebuah data, mengecek bidang yang kosong, kode yang hilang maupun mengindetifikasi data duplikat juga termasuk kedalam proses cleaning data.

2.5.3 Transformasi

Tujuan utama dalam transformasi data adalah mengubah data dengan tujuan mengubah skala data ke dalam bentuk lain agar memenuhi asumsi analisis. Data yang di tampilkan pada laporan merupakan data asli, data yang telah di transformasi hanya untuk membantu agar data asli dapat memenuhi asusmsi analisis. Data tersebut dapat diperkaya dengan dengan data maupun transpormasi eksternal yang relevan dengan tahapan KDD. Adapun teknik-teknik yang biasa digunakan dalam proses trandformasi data yakni *normalization*, pemilihan atribut dan *descritization*.

2.5.4 Data Mining

Data mining merupakan proses untuk menemukan pola ataupun informasi yang sebelumnya tidak dikeketahui dengan menggunakan suatu metode tertentu. Teknik, metode dan juga algoritma yang dapat digunakan sangat bervariasi. Penggunaan metode ataupun algoritma dapat dipilih tergantung kepada tujuan dan proses keseluruhan data mining. Pada proses ini seringkali memanfaatkan beberapa metode seperti matematika, statistika dan pemanfaatan kecerdasan buatan.

2.5.5 Interprestation atau Evaluation

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah di mengerti oleh banyak pihak yang memiliki kepentingan. Tahap ini memeriksa pola informasi yang bertentangan dengan fakta atau hipotesis yang ada pada sebelumnya.

2.6 Clustering

Clustering adalah metode khusus untuk pengumpulan data. *Clustering* adalah teknik pengumpulan data yang digunakan untuk menganalisis kelompok objek yang dikenal sebagai *cluster* yang dibuat dengan mengelompokkan objek yang lebih kecil berdasarkan kesamaan satu sama lain. Kemiripan yang menjadi dasar pengambilan keputusan tidak bersifat universal, sehingga dimensinya harus digali lebih jauh oleh peneliti atau analis. Diantara metode-metode untuk *data mining* atau analisis data, *clustering* merupakan teknik agregasi data yang sering digunakan. Proses menempatkan satu set objek data tertentu ke dalam kelompok yang disebut sebagai *cluster* disebut *clustering*. Oleh karena itu, metode pengelompokan berguna untuk menemukan kelompok yang tidak teridentifikasi dalam data.

2.7 DBSCAN

Algoritma Density-Based Spatial Clustering of Application with Noise (DBSCAN) merupakan algoritma berbasis kepadatan yang memiliki dua keunggulan. Keunggulan pertama yaitu dapat digunakan dalam dataset apa pun terlepas dari

bentuk distribusi data. Keuntungan kedua yaitu dapat diterapkan untuk menghilangkan *noise points* yang tersebar di dataset (Devi, AS et al, 2015). DBSCAN digunakan untuk mengelompokkan pengguna sehingga teridentifikasi pengguna yang memiliki karakteristik serupa (Min & Shuzhen, 2015). Algoritma DBSCAN di rancang untuk menemukan *cluster* dan *noise* yang terdapat pada dataset. Konsep yang mendasari dari proses algoritma ini ialah sebuah lingkungan radius (Eps) yang diberikan harus mengandung minimal jumlah titik (MinPts), yakni lingkungan harus melebihi beberapa *threshold* (persyaratan minimal)

Adapun langkah-langkah DBSCAN adalah sebagai berikut :

1. Inisialisasi parameter minpts, eps.
2. Tentukan titik awal atau p secara acak.
3. Ulangi langkah 3 – 5 hingga semua titik diproses.
4. Hitung eps atau semua jarak titik yang density reachable terhadap p.
5. Jika titik yang memenuhi eps lebih dari minpts maka titik p adalah core point dan *cluster* terbentuk.
6. Jika p adalah border point dan tidak ada titik yang *density reachable* terhadap p, maka proses dilanjutkan ke titik yang lain.

2.8 K-Means

K-means merupakan algoritma *cluster* yang banyak digunakan oleh peneliti, karena merupakan algoritma yang sederhana dan prosesnya cepat (Min & Shuzhen, 2015). Algoritma K-Means bekerja dengan cara mengelompokkan objek berdasarkan titik pusat *cluster* (*centroid*) terdekat dengan objek. Tujuannya adalah mengelompokkan objek dengan memaksimalkan kemiripan objek dalam satu *cluster* dan meminimalkan kemiripan objek antar *cluster.*

Ukuran kemiripan yang digunakan dalam *cluster* adalah fungsi jarak. Sehingga kemiripan objek dihitung berdasarkan jarak terpendek antara objek terhadap titik centroid. Salah satu metode yang digunakan untuk menghitung jarak *Euclidean Distance Space* dengan mengetahui jarak terpendek antara dua titik.

Dengan rumus sebagai berikut :

$$D(x_{i,}C_i) = \sqrt{\sum n}((x_{i,}C_i)^2 \quad \text{.....(1)}$$

Sesuai dengan rumus tersebut algoritma K-Means diawali dengan menentukan jumlah *cluster* (k) dan dilanjutkan dengan menentukan *centroid* (cj) setiap *cluster* secara acak. Langkah selanjutnya adalah menghitung jarak setiap objek (xi) ke *centroid* (cj) yang dinotasikan dengan D(xi,cj).

2.9 COSINE SIMILARITY

Cosine similarity digunakan dalam penelitian ini untuk menghitung kedekatan antar data. Pada umumnya, perhitungan pada metode *Cosine similarity* dihitung berdasarkan ukuran pada ruang vector. *Cosine similarity* menghitung kemiripan- kemiripan yang ada pada sebuah objek yang di representasikan sebagai dua vector. Berikut merupakan rumus dari *cosine similarity*

$$\cos(\theta) = \frac{\sum_k (d_{ik}d_{jk})}{\sqrt{\sum_k d_{ik}^2}\sqrt{\sum_k d_{jk}^2}} \text{.....(2)}$$

2.10 Rapid Miner

Pada penelitian ini, penulis menggunakan applikasi Rapid Miner. Applikasi ini dipilih sebab Rapid Miner dapat menyertakan visualisasi, statistik, dan informasi terpenting tentang model ke dalam laporan dengan sekali klik. Rapid Miner menyertakan pelaporan cerdas di mana pengguna dapat mengakses riwayat alur kerja untuk setiap *widget* dan visualisasi langsung dari laporan. Bukan hanya itu, applikasi ini memiliki antar muka yang baik, sehingga pengguna hanya *focus* pada analisis data bukan pada pengkodean. Membuat konstruksi *pipeline* analisis data yang kompleks menjadi sederhana. Rapid miner merupakan platform perangkat lunak data ilmu yang menyediakan lingkungan terpadu meliputi *mechine learning, deep learning, text mining* dan masih banyak lagi. Aplikasi ini mendukung semua langkah-langkah dalam proses data mining, seperti penyaiapan data, visualisasi data dan juga pengoptimalan data.