BAB 2
PENELITIAN TERKAIT DAN LANDASAN TEORI

2.1 Penelitian Terkait

Ada beberapa penelitian terkait dengan penggunaan *Data Mining* metode *cluster* dengan menggunakan Algoritma *Fuzzy C-Means* untuk dapat mengelompokkan bidang kerja berdasarkan lulusan, diantaranya adalah:

Penelitian oleh Cary Lineker Simbolon pada tahun 2013 [6]. Penelitian ini membahas tentang pengelompokkan lulusan jurusan matematika FMIPA Universitas Tanjungpura (UNTAN) yang membagi lulusan kedalam empat *cluster* berdasarkan IPK dan lama studi. Dari keempat cluster yang dihasilkan, cluster ke empat memiliki lulusan paling banyak yaitu 33 lulusan. *Cluster* keempat terdiri dari lulusan dengan kisaran lama studi 5,91 tahun. Dari hasil tersebut menunjukkan bahwa masih banyak mahasiswa jurusan Matematika di Fakultas MIPA Untan Pontianak yang menempuh lama studi lebih dari 10 smester atau 5 tahun. Sehingga hasil tersebut dapat dijadikan sebagai bahan pertimbangan jurusan dalam meningkatkan IPK mahasiswa untuk menyelesaikan masa studinya.

Berikutya penelitian oleh Sudirman pada tahun 2014 [7]. Penelitian ini membahas tentang penggunaan algoritma FCM yang diimplementasikan pada data status gizi balita di Puskesmas Kecamatan Belakang Padang, selanjutnya digunakan untuk mencari kesamaan terhadap perhitungan berdasarkan Standar Kementrian. Penelitian ini dilakukan sebab dalam pengolahan data pada Puskesmas ini masih menggunakan data arsip dan analisis belum tentu terhitung dengan baik. Akibat yang terjadi waktu akan lebih banyak terbuang dan dari segi hasil perhitungan juga belum tentu akurat. Maka, tentu diperlukan waktu tambahan guna mengoptimalkan data-data status gizi balita tersebut, sehingga dalam penelitian ini peneliti mencoba membangun aplikasi yang berbasis android untuk menyelesaikan masalah penentuan klasifikasi dengan menggunakan dua

perhitungan , yaitu berdasarkan Standar Kementerian RI Tahun 2010 tentang Standar Antropometri Penilaian Status Gizi Anak pada indeks Berat Badan per Tinggi Badan (BB/TB) dan perhitungan metode algoritma *Fuzzy C-Means*. Selanjutnya untuk penghitungannya variable yang digunakan dalam menentukan status gizi balita untuk kedua penghitungan tadi diantaranya adalah tinggi badan, berat badan, dan jenis kelamin. Dari 114 data sampel, kesamaan data yang telah diolah menggunakan algoritma *Fuzzy C-Means* menghasilkan jumlah kesamaan hasil klasifikasi terhadap perhitungan berdasarkan Standar Kementrian berjumlah 26 sampai dengan 32 data sampel. Kesamaan hasil klasifikasi yang dihasilkan system berkisar 22,81% hingga 28,07%. Dari hasil tersebut dapat disimpulkan bahwa aplikasi android mampu mengolah data balita serta menentukan klasifikasi untuk perhitungan menurut Standar Kementrian RI tahun 2010 dan untuk perhitungan menggunakan metode *Fuzzy C-Means*.

Dari penelitian terkait diatas dapat dirangkumkan pada table dibawah ini:

Tabel 2.1 Penelitian Terkait

No	Nama Peneliti dan Tahun	Masalah	Solusi	Hasil
1	Cary Lineker Simbolon, 2013	Bagaimana cara kerja algoritma *Fuzzy C-Means* untuk menyelesaikan masalah lulusan mahasiswa Matematika FMIPA UNTAN Pontianak.	Pembagian lulusan kedalam empat *cluster*, dimana *cluster* tersebut berdasarkan pada IPK dan lama studi mahasiswa.	Hasil dari pembahasan diketahui bahwa *cluster* ke empat memiliki anggota lulusan paling banyak dengan kisaran lama studi 5,91 tahun.
2	Sudirman, 2014	System analisa status gizi dilakukan secara manual,	Pembuatan aplikasi berbasis android untuk	Terdapat kemiripan sebanyak 26 sampai 32 dari 114 sampel

No	Nama Peneliti dan Tahun	Masalah	Solusi	Hasil
		mengakibatkan banyak waktu terbuang dan perhitungan yang tidak akurat.	menyelesaikan masalah penentuan klasifikasi dengan menggunakan dua perhitungan, yaitu berdasarkan Standar Kementerian (SK) RI Tahun 2010 dan FCM.	berdasarkan perhitungan menggunakan SK dan FCM, sedangkan dalam presentase system terdapat kemiripan sebesar 22,81 % sampai 28,07 %.

Dari kedua penelitian tersebut, terdapat perbedaan penelitian dimana penelitian pertama yang dilakukan oleh Cary Lineker Simbolon dengan menggunakan algoritma *Fuzzy C-Means* untuk menyelesaikan masalah lulusan mahasiswa Matematika FMIPA pada UNTAN Pontianak dengan membagi lulusan kedalam empat *cluster*, dimana *cluster* tersebut berdasarkan pada IPK dan lama studi mahasiswa. Dari penelitian yang telah dilakukan menghasilkan lulusan terlama yaitu 5,91 tahun terdapat pada *cluster* ke empat. Sedangkan pada penelitian yang dilakukan oleh Sudirman yaitu pemanfaatan algoritma *Fuzzy C-Means* untuk mengurangi waktu yang terbuang pada analisa status gizi dan penghitungannya yang tidak akurat. Metode yang digunakan yaitu dengan cara membandingkan algoritma *Fuzzy C-Means* dengan perhitungan berdasarkan Standar Kementerian (SK) RI Tahun 2010 yang menghasilkan kemiripan sebesar 22,8% sampai dengan 28,07%. Berdasarkan dari kedua penelitian tersebut, penulis tertarik untuk menggunakan metode *clustering* dengan algoritma *Fuzzy C-Means* pada penelitian ini, yakni pada Universitas Dian Nuswantoro dengan memanfaatkan data transkip nilai mahasiswa guna mencari kesesuaian antara kemampuan

akademis dan bidang pekerjaan yang selanjutnya mengelompokkan bidang pekerjaan berdasarkan nilai mata kuliah mahasiswa.

2.2 Landasan Teori

2.2.1 Bidang Pekerjaan

Menurut Kamus Besar Bahasa Indonesia (KBBI) bidang pekerjaan terdiri dari dua kata yaitu bidang dan pekerjaan. Bidang merupakan penggolongan bagi sesuatu yang luas, sedangkan pekerjaan itu sendiri memiliki arti sesuatu yang dilakukan untuk mendapat nafkah [8]. Apabila digabungkan kedua kata tersebut memiliki arti yaitu penggolongan pekerjaan sesuai keahlian tenaga kerja.

2.2.2 Lulusan

Dilihat dari definisi menurut KBBI, lulusan berarti yang sudah lulus dari ujian [8]. Apabila dilihat dari konteks penelitian ini lulusan yang dimaksud adalah mahasiswa yang sudah melewati serangkaian ujian yang diberikan oleh pihak universitas dan telah mendapatkan gelar dari pendidikan strata satu (S-1).

2.2.3 *Data Mining*

Istilah data mining sebenarnya mulai dikenal sejak tahun 1990, pada saat itu pemanfaatan data sangatlah penting dalam berbagai bidang seperti akademik, bisnis, hingga medis. Data mining dapat diterapkan pada berbagai bidang yang memiliki sejumlah data, namun karena wilayah penelitian beserta sejarah yang belum lama, maka data mining masih mengalami perdebatan mengenai posisinya dibidang pengetahuan. Sehingga dari sini, Darly Pregibon menyatakan bahwa data mining merupakan perpaduan antara statistik, kecerdasan buatan, dan riset basis data [9].

Apabila dilacak, data mining memiliki empat akar ilmu diantaranya adalah [9]:

1. Statistik

Akar yang paling tua diantara lainnya, mungkin tanpa adanya statistik data mining juga tidak akan pernah ada. Dengan menggunakan statistik klasik data dapat diringkas kedalam *exploratory data analysis* (EDA). EDA berguna untuk mengidentifikasi hubungan sistematis antar variable ketika tidak ada cukup informasi alami yang dibawanya.

2. Kecerdasan Buatan atau *artificial intelligence* (AI)

   Bidang ilmu ini berbeda dari bidang sebelumnya. Teorinya dibangun berdasarkan teknik heuristik sehingga AI berkontribusi terhadap teknik pengolahan informasi berdasarkan pada model penalaran manusia.

3. Pengenalan Pola

   Data mining juga meropakan turunan dari bidang pengenalan pola,tapi hanya pada pengolahan data dari basis data. Data yang diambil dari basis data kemudian diolah bukan dalam bentuk relasi, namun diolah dalam bentuk normal pertama biarpun begitu data mining memiliki ciri khas yaitu pencarian pola asosiasi dan pola sekuensial.

4. Sistem Basis Data

   Akar bidang ilmu yang terakhir dari data mining yang menyediakan informasi berupa data yang akan ‘digali’ menggunakan metode-metode yang ada pada data mining.

Setelah mengulas sejarahnya, *Data mining* sendiri merupakan suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. Dilihat dari sudut pandang bahasa, *Data Mining* diartikan sebagai penambangan data. *Data Mining* merupakan analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya. Terdapat pendapat lain yang menyatakan bahwa *Data Mining* merupakan analisis dari sekumpulan data yang diamati untuk menemukan hubungan yang tidak terduga dan merangkum data dengan cara yang baru yang dapat dipahami dan berguna bagi pemilik data [10]. Beberapa teori juga menyebutkan bahwa data mining merupakan sebuah proses

yang mempekerjakan satu atau lebih teknik pembelajaran computer untuk menganalisis dan mengekstraksi pengetahuan secara otomatis [11].

Walaupun data mining diartikan sebagai penemuan informasi, tidak semuanya disebut sebagai data mining. Berikut beberapa contoh yang membedakannya:

1. Bukan data mining: Pencarian informasi tertentu di internet.
   Data mining: Pengelompokkan informasi yang mirip dalam konteks tertentu pada hasil pencarian.
2. Bukan data mining: Petugas medis mencari data medis untuk menganalisis catatan pasien dengan penyakit tertentu.
   Data mining: Peneliti medis mencari cara pengelompokan data penyakit pasien berdasarkan data diagnosis, umur, dan alamat.
3. Bukan data mining: Analisis gambar laporan keuangan penjualan perusahaan.
   Data nining: Menggunakan basis data transaksi perusahaan dengan fokus pada data sales untuk mengidentifikasi profil utama pelanggan.
4. Bukan data mining: Pembuatan laporan penjualan tahunan dengan merekap data selama setahun.
   Data mining: Memanfaatkan data penjualan perusahaan untuk mendapatkan pola prediksi stok yang sebaiknya disediakan untuk tahun berikutnya.

Beberapa tantangan dalam data mining diantaranya adalah [11]:

1. *Scalability*, merupakan besarnya ukuran data yang digunakan.
2. *Dimentionality*, merupakan banyaknya jumlah atribut dalam data yang nantinya akan diproses.
3. *Complex and Heterogeneus Data*, merupakan data yang sifatnya komplek dan memiliki variasi yang beragam.
4. *Data Quality*, merupakan kualitas data yang nantinya akan diproses misalnya data yang bersih dari *noise, missing value,* dsb.
5. *Data Ownership and Distribution,* merupakan siapa pemilik data dan bagaimana distribusinya.

6. *Privasi Preservation,* merupakan kerahasiaan data yang banyak diterapkan pada data nasabah perbankan.
7. *Streaming Data,* merupakan aliran data itu sendiri.

2.2.4 *Clustering*

Ada saat dimana set data yang akan diproses dalam data mining belum diketahui label kelasnya. Misalnya terdapat kasus data catatan akademik mahasiswa, diketahui jumlah dari SKS yang sudah dilalui dan jumlah IPK yang didapat. Sebelum dilakukannya pemrosesan menggunakan data mining, belum diketahui label dari kelompok mahasiswa tersebut. Selanjutnya ketika telah dilakukan pemrosesan dengan menerapkan algoritma yang telah ditentukan lalu data akan diproses oleh algoritma untuk selanjutnya dikelompokkan berdasarkan karakter alaminya. Data yang sudah diproses menggunakan algoritma yang sudah ditentukan akan berkelompok dengan sendirinya. Data yang mirip dengan data lain akan berkelompok kedalam satu cluster, dan yang tidak sesuai akan membuat klompok sendiri dengan data yang lainnya. Misalnya data yang sudah didapat akan diklompokkan kedalam tiga klompok yaitu klompok pertama data dengan SKS sedikit dan IPK tinggi, klompok data berikutnya yaitu klompok data dengan SKS tinggi dan IPK rendah, dan kelompok yang terakhir yaitu klompok dengan data SKS rendah dan IPK rendah. Permasalahan pengelompokkan data tersebut disebut dengan pembelajaran tidak terbimbing atau *unsupervised learning* [9].

Clustering sendiri merupakan proses mengelompokkan suatu set obyek menjadi kelas-kelas yang terdiri dari obyek yang sama. Lebih sederhananya *clustering* adalah sebuah proses mengelompokkan obyek berdasarkan kesamaan karakteristik diantara obyek-obyek tersebut [5]. Teknik ini banyak sekali diterapkan diberbagai bidang seperti kesehatan, psikologi, klimatologi, statistik dan sebagainya.

Clustering data dapat dibedakan menjadi dua tujuan, yaitu untuk pemahaman dan penggunaan [9]. Bila tujuannya untuk pemahaman maka cluster yang terbentuk harus menangkap struktur alami data. Sedangkan apabila tujuannya untuk

penggunaan biasanya tujuan utamanya untuk mencari prototipe cluster yang paling representatif terhadap data sehingga memberikan abstraksi dari setiap objek data yang berada di dalam cluster. contoh-contoh tujuan clustering sebagai pemahaman sebagai berikut [9]:

1. Biologi

   Hewan-hewan yang berada di alam sudah banyak sekali yang diketahui, dan hewan-hewan tersebut diklompokkan menurut karakter tertentu secara hierarki yaitu: kerajaan, filum, kelas, ordo, family, genus, dan spesies. Dari kelompok tersebut level tertinggi adalah kerajaan dan yang terendah adalah spesies. Satu jenis hewan memiliki nama spesies sendiri, dua hewan dengan spesies berbeda bisa memiliki kelas genus yang sama, sejumlah hewan yang berbeda bisa memiliki family yang sama, begitu pula di level ordo, kelas, filum dan kerajaan. Semua hewan berada didalam kelompok hewan yang sama di level kerajaan yaitu hewan. Teknik clustering dalam bidang ini yang lain seperti pengelompokkan gen-gen yang memiliki fungsi yang sama.
2. Pencarian Informasi

   Website yang terdapat di internet berjumlah jutaan bahkan miliaran sehingga ketika dilakukan pencarian, mesin pencari akan memberikan hasil ribuan halaman. Disini teknik clustering dapat digunakan sebagai pembantu dalam mengelompokkan hasil halaman yang diberikan mesin pencari kedalam jumlah kecil dimana setiap kelompok berisikan halaman yang memiliki karakteristik yang sama mirip.
3. Klimatologi

   Analisis cluster disini dapat digunakan untuk menemukan pola pada tekanan udara diwilayah kutub dan lautan yang memiliki pengaruh besar terhadap cuaca yang terjadi di daratan.
4. Bisnis

   Dibidang ini clustering digunakan untuk mensegmentasi pelanggan dalam kelompok-kelompok kecil yang bertujuan untuk analisis dan strategi marketing dengan memanfaatkan data pelanggan yang ada.

5. *Summarization*

   Banyaknya data berpengaruh pada biaya ketika dilakukannya summarization, karena semakin banyak data maka akan semakin mahal biayanya. Solusi untuk masalah ini yaitu diterapkannya teknik clustering untuk membuat prototipe yang dapat mewakili kondisi seluruh data, misalnya mengambil nilai rata-rata bagi semua data dari setiap cluster yang akan diwakili sehingga sejumlah data yang tergabung dalam cluster hanya akan diwakili sebuah data. Dengan cara seperti itu maka waktu dan kompleksitas komputasi dapat dikurangi secara signifikan.

6. Kompresi

   Data yang tergabung didalam tiap-tiap cluster dapat dianggap memiliki karakteristik yang sama, sehingga data dalam cluster yang sama dapat dikompresi dengan diwakili oleh indeks prototyipe yang dikaitkan dengan sebuah cluster, teknik seperti ini lebih dikenal sebagai teknik *quantization*.

7. Pencarian tetangga terdekat secara efisien

   Pada teknik K-NN, komputasi yang digunakan sebagai pencarian tetangga terdekat semakin berat apabila data yang diproses semakin banyak, padahal hal tersebut tidak sebanding dengan jumlah data yang nantinya digunakan sebagai tetangga terdekat. Dengan cara ini komputasi pencarian tetangga terdekat dapat digantikan prototipe terdekat sehingga mampu mengurangi waktu komputasi secara signifikan.

Berikutnya menurut struktur, clustering dibedakan menjadi dua, yaitu *hierarki* dan *partisi* [9]. Pada pengelompokan berbasis hierarki semua data dapat bergabung menjadi sebuah cluster. Sedangkan pada pengelompokan berbasis partisi akan membagi setiap data hanya menjadi anggota satu cluster saja.

Pembagian berdasarkan keanggotaan data dalam cluster dapat dibedakan menjadi dua, yaitu *eksklusif* dan *tumpang-tindih* [9]. Termasuk kategori eksklusif apabila sebuah data hanya menjadi satu anggota cluster. beberapa metode clustering yang termasuk dalam kategori ini diantaranya *K-Means*, *DBSCAN*, dan *Self Organizing Map* (SOM). Berikutnya yang termasuk dalam kategori tumpang-tindih yaitu metode yang membolehkan sebuah data menjadi anggota lebih dari satu cluster misalnya *Fuzzy C-Means* dan pengelompokan yang berbasih *hierarki*.

Sementara menurut kategori kekompakan, clustering dibagi menjadi menjadi dua bagian yaitu, *komplet* dan *parsial*. Sebuah data dikatakan kompak menjadi satu cluster apabila data tersebut dapat bergabung menjadi satu, dan dikatakan memiliki perilaku menyimpang apabila ada beberapa data yang tidak ikut bergabung ke dalam cluster mayoritas. Data tersebut dikenal sebagai *outlier*, *noise*, atau bahkan *uninterested background.*

2.2.5 FCM *(Fuzzy C-Means)*

Metode *Fuzzy C-Means* (FCM) didasakan pada teori logika fuzzy yang diperkenalkan pertamakali oleh Lotfi Zadeh [9]. FCM merupakan suatu teknik pengclusteran data yang keberadaan tiap-tiap titik data suatu cluster ditentukan oleh nilai keanggotaan. Nilai keanggotaan tersebut akan mencakup bilangan real pada interval 0-1 [12]. Tujuan dari algoritma FCM yaitu untuk mendapatkan pusat cluster yang nantinya akan digunakan untuk mengetahui data yang masuk kedalam cluster. Berikut adalah penjabaran dari algoritma FCM [6]:

1. Menentukan data yang akan dicluster *X*, berupa matriks berukuran *n* x *m* (*n*=jumlah sampel data, m=atribut setiap data). X_{ij}=data sampel ke-i (i=1,2,...,*n*), atribut ke-j (j=1,2,...,m).
2. Tentukan jumlah cluster (c), pangkat (w), maksimum iterasi (MaxIter), error terkecil yang diharapkan (ζ), fungsi obyektif awal (P0=0), iterasi awal (t=1).
3. Bangkitkan bilangan random μ_{ik}, i=1,2,...,n; k=1,2,...,c; sebagai elemen-elemen matriks partisi awal U. Matriks partisi (U) pada pengelompokan fuzzy memenuhi kondisi sebagai berikut [6]:

$$\mu_{\text{ik}}\varepsilon[0,1]; \qquad 1 \leq i \leq n; \qquad 1 \leq k \leq c \qquad \textbf{2.1}$$

μ_{ik} adalah derajat keanggotaan yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota ke dalam suatu cluster. Hitung jumlah setiap kolom (atribut):

$$Q_i = \sum_{k=1}^{c} \mu_{ik} \qquad \textbf{2.2}$$

$Q_i = \mu_{i1} + \mu_{i2} + \cdots + \mu_{ic}$

dengan $i = 1,2, \ldots, n$

4. Hitung pusat cluster ke-k: Vkj, dengan k=1,2,...,c; dan j=1,2,...,m

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik})^{w} X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}} \quad \textbf{2.3}$$

5. Hitung fungsi obyektif pada iterasi ke-t, Pt:

   Fungsi obyektif digunakan sebagai syarat perulangan untuk mendapatkan pusat cluster yan tepat. Sehingga diperoleh kecenderungan data untuk masuk ke cluster mana pada step akhir. Untuk iterasi awal nilai t=1.

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right](\mu_{ik})^w\right) \quad \textbf{2.4}$$

6. Hitung perubahan matriks partisi:

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}} \quad \textbf{2.5}$$

7. Cek kondisi berhenti:

   a. |Pt – Pt-1| < ζ) atau (t>MaxIter) maka berhenti;

   b. Jika tidak, iterasi dinaikkan t=t+1, ulangi langkah ke-4.