LAPORAN PENELITIAN

Analisis Cluster Dengan Menggunakan Metode Hierarchical Berdasarkan Pengelompokan Pengeluaran Kacang-Kacangan Kabupaten/Kota Tahun 2021-2022

Dosen Pengampu: Dr. Ir. Ananto Tri Sasongko, M.Sc.

Oleh Kelompok 6:

1. **Alvina Damayanti (312110125)**
2. **Laela Nur Rohmah (312110425)**
3. **Nurul Najwa Sabilla (312110451)**
4. **Sara Khusnul Mumtazah (312110319)**

FAKULTAS TEKNIK

TEKNIK INFORMATIKA

UNIVERSITAS PELITA BANGSA

2024

DAFTAR ISI

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tanaman kacang-kacangan sudah ditanam di Indonesia sejak dulu. Tanaman ini terdiri atas berbagai jenis, misalnya kacang kedelai, kacang hijau, kacang tanah, dan berbagai jenis kacang sayur misalnya kecipir, kapri, kacang panjang dan buncis. Perhatian pemerintah terhadap tanaman kacang-kacangan sangat besar. Dalam Repelita VI, pemerintah memprogramkan pembangunan subsektor pertanian tanaman pangan dan hortikultura termasuk palawija, terutama kacang-kacangan. Permintaan terhadap kacang-kacangan pada masa yang akan datang, diperkirakan akan terus meningkat seiring dengan pertambahan jumlah penduduk. Mengacu pada Pola Pangan Harapan (PPH) tahun 2000, konsumsi rata-rata kacang-kacangan penduduk Indonesia adalah sebesar 35,88 gr/kapita/hari (Astawan, 2009).

Sebagaimana kita ketahui bahwa kacang-kacangan di Indonesia memiliki beragam jenis kacang-kacangan yang dapat tumbuh dengan baik. Beberapa kacang lokal dapat ditemui di daerah dan digunakan untuk kebutuhan pangan. Selain dapat langsung dikonsumsi, kacang-kacangan dapat diolah menjadi beberapa produk makanan, antara lain kacang telur, selai kacang, tempe, tahu, dan berbagai olahan lainnya. Kacang yang telah diolah dapat menambah nilai ekonomis dan nilai unggul. Mengingat komoditi ini memiliki potensial sebagai sumber zat gizi lain, yaitu mineral, vitamin B, karbohidrat kompleks dan serat makanan. Kandungan seratnya tinggi, maka kacang-kacangan dipilih untuk menjadi sumber serat. Kacang-kacangan memberikan sekitar 135 kkal per 100 gram bagian yang dapat dimakan. Konsumsi kacang-kacangan sebanyak 100 gram, maka jumlah itu akan mencukupi sekitar 20% kebutuhan protein dan 20% kebutuhan serat per hari. Kacang-kacangan merupakan sumber protein yang baik, dengan kandungan protein berkisar antara 20 – 30%, selain sumber protein juga mengandung senyawa lainnya seperti mineral, vitamin B1, B2, B3, karbohidrat, dan serat (Koswara, 2009).

Survei Sosial Ekonomi Nasional (Susenas) yang dilakukan Badan Pusat Statistik (BPS), penduduk Indonesia di perkotaan dan perdesaan rata-rata memiliki pengeluaran untuk makanan sebesar Rp.631,600 ribu/kapita/bulan. Namun, terlihat dari pengeluaran konsumsi per kapita penduduk Indonesia untuk kelompok komoditas makanan per September 2021, kacang-kacangan yaitu Rp.14.480/kapita/bulan. Dibandingkan dengan pengeluaran untuk protein hewani memiliki rentang pengeluaran konsumsi dari Rp29.907 sampai dengan Rp. 53.118. Selain itu, dalam beberapa tahun terakhir, semakin banyak orang yang mengganti protein nabati dengan protein hewani, sehingga menurunkan permintaan terhadap kacang-kacangan, yang merupakan sumber utama protein nabati (Singh, Jain, Ujinwal, dan Lang Yan, 2022).

Dalam hal menggali informasi tidak dapat dilakukan dengan mudah. Dengan begitu membutuhkan bantuan teknologi *data mining.* Teknologi *data mining* adalah salah satu alat untuk mengekstraksi data dalam database besar dan dengan spesifikasi kerumitan yang tinggi. Terdapat algoritma-algoritma yang memiliki peran penting dalam mewujudkan nilai nyata dari kumpulan data yang besar dan kompleks.

Penelitian ini akan menerapkan metode *clustering*. *Clustering* merupakan pendekatan menggunakan algoritma untuk menganalisis dan mengidentifikasi pola dari suatu dataset tanpa

adanya bantuan atau intervensi langsung dari manusia. Dalam konteks ini, penelitian tidak menentukan informasi atau keluaran yang diharapkan dari suatu input terhadap algoritma. Sebaliknya, algoritma diberikan kebebasan untuk mengeksplorasi dan mengidentifikasi pola yang mungkin ada dalam dataset tanpa panduan manusia. Algoritma yang akan digunakan dalam penelitian ini akan diimplementasikan menggunakan pemrograman Python untuk memfasilitasi proses analisis dan eksplorasi data yang efisien dan fleksibel.

Hierarchical clustering adalah salah satu metode dalam analisis cluster yang digunakan untuk mengelompokkan data ke dalam struktur *hierarchical* atau pohon. Dalam *hierarchical clustering*, setiap data awal dianggap sebagai satu cluster, dan kemudian cluster-cluster tersebut digabungkan secara berurutan berdasarkan kemiripan atau kedekatan antara elemen-elemen di dalamnya. Proses ini terus berlanjut hingga semua data tergabung dalam satu cluster besar atau beberapa kluster yang tingkatannya saling berbeda. Pendekatan umum dalam *hierarchical clustering* adalah *agglomerative* (*bottom-up*).

Berdasarkan latar belakang diatas maka penulis melibatkan pengelompokan data. Mengetahui analisis *data mining* untuk mengelompokkan wilayah-wilayah berdasarkan rata-rata pengeluaran perkapita seminggu pada kacang-kacangan dan kemudian memberikan pemahaman tentang pola konsumsi atau struktur ekonomi yang mungkin terkait dengan setiap kelompok. Serta, mengidentifikasi pola konsumsi kacang yang umum di antara kelompok konsumen tertentu pada tahun 2021 sampai 2022. Penulis menggunakan metode analisis *hierarchical clustering* yaitu metode *average*, dimana pada metode ini proses pengelompokan dimulai dengan menghitung rata-rata antara dua objek. Kelebihan metode ini adalah dapat menggabungkan objek kedalam cluster dengan ragam yang kecil, memperhatikan struktur cluster yang terbentuk serta lebih stabil.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang penelitian ini, perumusan masalah yang diambil sebagai berikut: “Bagaimana menerapkan konsep *data mining* melalui proses algoritma *hierarchical clustering* metode *average* untuk mengelompokkan wilayah-wilayah berdasarkan rata-rata pengeluaran perkapita seminggu pada kacang-kacangan dan kemudian memberikan pemahaman tentang pola konsumsi atau struktur ekonomi yang mungkin terkait dengan setiap kelompok. Serta, mengidentifikasi pola konsumsi kacang yang umum di antara kelompok konsumen tertentu pada tahun 2021 sampai 2022”.

1.3 Batasan Masalah

Keterbatasan masalah untuk analisis ini meliputi:

a. Analisis ini mengambil dataset dari Badan Pusat Statistika (BPS) tentang Konsumsi dan Pendapatan
b. Data yang digunakan merupakan data mengenai Rata-rata Pengeluaran Perkapita Seminggu Menurut Kelompok Kacang-Kacangan Per Kabupaten/kota (Rupiah/Kapita/Minggu) Tahun 2021-2022
c. Variabel masukan yang digunakan adalah kacang-kacangan, kacang tanah tanpa kulit, kacang kedelai, kacang lainnya, tahu, tempe, dan oncom

d. Output yang akan dihasilkan adalah untuk mengelompokkan wilayah-wilayah berdasarkan rata-rata pengeluaran perkapita seminggu pada kacang-kacangan dan kemudian memberikan pemahaman tentang pola konsumsi atau struktur ekonomi yang mungkin terkait dengan setiap kelompok. Serta, mengidentifikasi pola konsumsi kacang yang umum di antara kelompok konsumen tertentu pada tahun 2021 sampai 2022.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah, tujuan dari analisis ini adalah menerapkan konsep *data mining* melalui proses algoritma *hierarchical clustering* metode *average*, untuk mengelompokkan wilayah-wilayah berdasarkan rata-rata pengeluaran perkapita seminggu pada kacang-kacangan dan kemudian memberikan pemahaman tentang pola konsumsi atau struktur ekonomi yang mungkin terkait dengan setiap kelompok. Serta, mengidentifikasi pola konsumsi kacang yang umum di antara kelompok konsumen tertentu pada tahun 2021 sampai 2022.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

a. Bagi penulis, analisis ini berguna untuk menambah wawasan tentang *data mining* melalui proses algoritma *hierarchical clustering* metode *average*
b. Bagi pembaca, diharapkan dapat dijadikan sebagai informasi baru dan sebagai bahan referensi serta sebagai tolak ukur bagi pembaca yang ingin melakukan analisis lebih lanjut mengenai permasalahan yang sama.

1.6 Sistematika Penulisan

Sistematik penelitian yang digunakan dalam laporan penelitian ini dibagi menjadi lima bab, yaitu:

a. BAB I PENDAHULUAN
   Bab ini berisi latar belakang, perumusan masalah, batasan masalah, manfaat penelitian, tujuan penelitian, dan sistematika penulisan.
b. BAB II TINJAUAN PUSTAKA
   Bab ini menjelaskan studi sebelumnya dan teori yang mendukung penelitian (landasan teori).
c. BAB III METODOLOGI PENELITIAN
   Bab ini berisi metode dan langkah yang digunakan dalam penelitian.
b. BAB IV ANALISA DAN PEMBAHASAN
   Bab ini berisi hasil dan diskusi dari penelitian yang dilakukan. Diskusi ini berisi analisis data, langkah penyelesaian, desain, dan analisis proses dan hasil.
c. BAB V KESIMPULAN DAN SARAN
   Bab ini berisi kesimpulan dari hasil penelitian yang telah dilakukan dan saran untuk pengembangkan penelitian lebih lanjut

BAB II

TINJAUAN PUSTAKA

2.1 Data Mining

Data mining adalah proses pencarian pola-pola yang tersembunyi (*hidden patern*) berupa pengetahuan (*knowledge*) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada dalam *database*, dan *dataware*, atau media penyimpanan informasi yang lain. *Data mining* menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin untuk mengekstrak dan mengidentifikasi informasi yang berguna dan pengetahuan terkait dari berbagai database besar . *Data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari data yang sangat besar. Menurut Pramudiono (2007*) data mining* juga dapat disebut sebagai rangkaian proses untuk mengeksplorasi nilai tambah dalam bentuk pengetahuan yang belum diketahui secara manual dari kumpulan data

2.1.1 Teknik Data Mining

Ada lima jenis Teknik analisis yang dapat dikalasifikasikan, yaitu:

a. Asosiasi

   Teknik asosiasi adalah teknik penambangan untuk menemukan aturan asosiatif antara kombinasi atribut. Contoh aturan asosiatif dari analisis pembelian di supermarket diketahui seberapa besar kemungkinan bagi pelanggan untuk membeli roti bersama dengan susu. Dengan pengetahuan ini, pemilik supermarket dapat mengatur penempatan barang-barang mereka atau merancang strategi pemasaran menggunakan kupon diskon untuk kombinasi barang tertentu.

b. Klasifikasi

   Klasifikasi adalah proses menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri dapat menjadi aturan "if-then", dalam bentuk pohon keputusan, rumus matematika atau jaringan saraf. Proses klasifikasi biasanya dibagi menjadi dua fase: pembelajaran dan pengujian.

c. Pengelompokan (*Clustering*)

   Tidak seperti association rule mining dan klasifikasi di mana kelas data telah ditentukan sebelumnya, pengelompokan melakukan pengelompokan data tanpa berdasarkan pada kelas data spesifik. Bahkan pengelompokan dapat digunakan untuk memberi label pada kelas data tidak dikenal. Oleh karena itu pengelompokan sering diklasifikasikan sebagai metode pembelajaran yang tidak diawasi. Prinsip pengelompokan adalah untuk memaksimalkan kesamaan antara anggota satu kelas dan meminimalkan kesamaan antar kelompok. *Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensional.

d. Estimasi
   Estimasi hampir sama dengan klasifikasi, kecuali variable taget estimasi lebih kearah numerik dari pada kearah kategori. Model dibangun dengan record lengkap menyediakan nilai dari variable target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variable target dibuat berdasarkan nilai variable prediksi.
e. Prediksi
   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

2.1.2 Tahap-Tahap Data Mining

Istilah *data mining* dan *Knowledge Discovery in Database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Walaupun sebenarnya *data mining* sendiri adalah bagian dari tahapan proses dalam KDD.

a. *Data cleaning*, untuk membersihkan data dari noise data dan data yang tidak konsisten. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan mempebaiki kesalahan pada data
b. *Data integration*, mengkombinasikan atau mengintegrasikan beberapa sumber data.
c. *Data selection*, mengambil data-data yang relevan dari database untuk dianalisis.
d. *Data transformation*, mentransformasikan data summary ataupun operasi agregasi.
e. *Data mining*, merupakan proses yang esensial dimana metode digunakan untuk mengekstrak pola data yang tersembunyi dengan menggunakan Teknik atau metode tetentu. Teknik, metode, atau algoritma dalam data mining sangat bevariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
f. *Interpretation / Evaluasi*, pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.2 Analisis Cluster

Analisis *cluster* merupakan salah satu analisis multivariat (banyak variabel) yang berfungsi untuk mengelompokkan objek-objek atau beberapa variabel berdasarkan karakteristik yang dimiliki. Selain itu, analisis *cluster* juga bertujuan dalam memaksimalkan kesamaan objek dalam *cluster* sementara itu juga memaksimalkan perbedaan antar *cluster* (Hair, 2009).

Analisis *cluster* bermanfaat dalam menyimpulkan suatu data yang kompleks dengan cara mengelompokkan objek-objek yang memiliki kemiripan karakteristik. Adapun *cluster* yang baik mempunyai ciri-ciri sebagai berikut (Santoso, 2015):

a. Memiliki kesamaan (*homogenitas*) yang tinggi antar objek dalam satu cluster (*within cluster*)

b. Memiliki perbedaan (heterogenitas) yang tinggi antar *cluster* yang satu dengan *cluster* yang lainnya (*between cluster*).

Berdasarkan ciri-ciri tersebut, maka dapat dinyatakan bahwa sebuah *cluster* yang efektif merupakan *cluster* yang terdiri dari beberapa objek yang memiliki kemiripan antara *cluster* satu dengan lainnya, namun sangat berbeda dengan *cluster* yang lain. Dalam hal ini, kata mirip juga dapat diartikan dengan tingkat kesamaan karakteristik antara dua objek (Santoso, 2015).

2.2.1 Menetapkan Ukuran Jarak Antar-Data

Analisis *cluster* membutuhkan beberapa ukuran untuk mengetahui kemiripan antara objek-objek yang akan diteliti. Ukuran yang biasa digunakan dalam mengukur kemiripan antar data pada analisis *cluster* adalah ukuran jarak (*distance*). Secara umum, diperoleh tiga ukuran dalam mengukur kemiripan antar data, yaitu asosiasi, korelasi, dan kedekatan (Sukmawati, 2017). Ukuran kedekatan yang digunakan untuk menghitung jarak antar cluster adalah sebagai berikut:

a. Jarak Euclidean (*Euclidean Distance*)

Jarak euclidean (*euclidean distance*) antara dua objek dapat terdefinisikan dengan spesifik. Jarak ini digunakan jika variabel tidak berkorelasi. Jarak euclidean (*euclidean distance*) didefinisikan:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(y_{ik} - y_{jk})^2; i, j = 1, 2, \dots, n}$$

Dimana:

d_{ij} = jarak antara objek ke-i dan ke-j
y_{ik} = nilai pengamatan objek ke-i dan ke-j
y_{jk} = nilai pengamatan objek ke-j dan ke-k
p = banyaknya variabel

b. Jarak Square Euclidian (*Squared Euclidean Distance*)

Jarak square euclidian (*squared euclidean distance*) adalah hasil variasi dari jarak euclidian (*euclidean distance*) dimana dalam jarak square euclidian (*squared euclidean distance*) akar tersebut dihapuskan.

$$d_{ij} = \sum_{k=1}^{p}(y_{ik} - y_{jk})^2; i, j = 1, 2, \dots, n$$

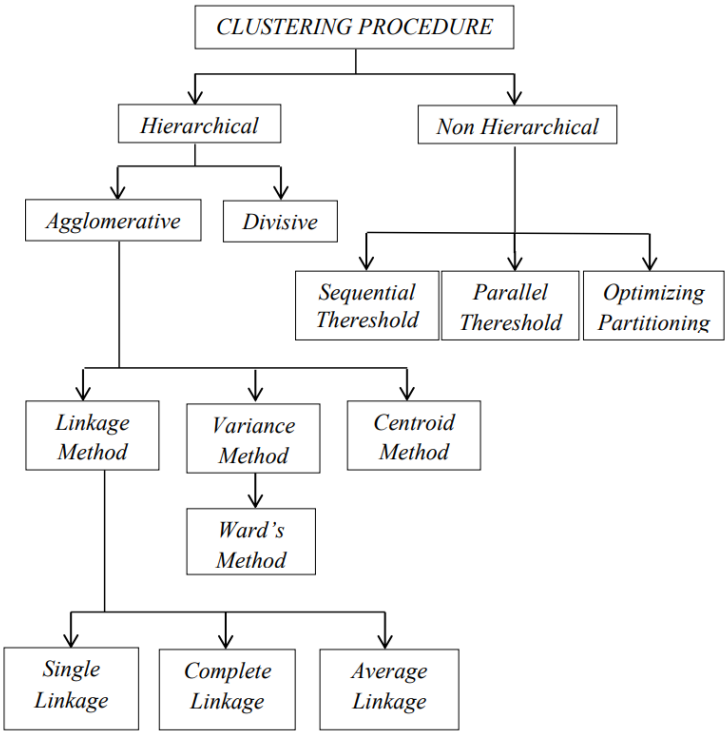
Dimana:

d_{ij} = jarak antara objek ke-i dan ke-j
y_{ik} = nilai pengamatan objek ke-i dan ke-j
y_{jk} = nilai pengamatan objek ke-j dan ke-k
p = banyaknya variabel

2.2.2 Melakukan Proses Clustering

Proses *clustering* merupakan proses yang dilakukan dengan dua metode, yaitu dengan metode hierarki dan non hierarki. Pada metode hierarki terdapat dua metode diantaranya metode penggabungan (*agglomerative*) dan metode pemisah (*devisive*). Pada metode penggabungan (*agglomerative*) merupakan cara dengan memandang setiap objek pengamatan berasal dari *cluster* yang berbeda. Secara bertahap objek yang memiliki kedekatan dengan objek lainnya akan digabungkan menjadi satu *cluster*. Kemudian tahap selanjutnya objek yang memiliki kedekatan dengan objek kedua. Hal ini berlangsung hingga semua objek berada dalam *cluster*. Sebaliknya pada metode pemisahan (devisive) merupakan cara dengan memandang semua objek berasal dari satu *cluster* yang sama. Kemudian menentukan perbedaan diantara objek tersebut. Objek yang berbeda akan dikeluarkan dari *cluster* pertama dan seterusnya secara bertahap hingga akan terbentuk cluster terakhir yang beranggotakan satu objek saja.

Metode non hierarki terdiri dari tiga metode, yaitu metode *sequential thereshold*, metode *parallel*, metode *optimizing partitioning*. Klasifikasi prosedur pengclusteran analisis *cluster* ini ditampilkan dalam bagan dibawah ini (Simamora, 2005).



Gambar 2.3.2. Bagan Analisis Cluster

1. Metode Hierarki (*Hierarchical Method*)

   Metode hierarki (*hierarchical method*) merupakan metode yang dilakukan secara bertahap. Pada metode ini akan membentuk tahapan tertentu seperti pada struktur pohon serta dapat di hasilkan bentuk dendogram. Dendogram merupakan epresentasi visual dari tahap-tahap proses analisis cluster yang terbentuk dan nilai koefiensi jarak pada setiap tahap. Hasil berupa angka yang berada disebelah kanan dendogram merupakan objek penelitian, karena terdapat garis yang menghungkan objek-objek

tersebut dengan objek-objek lainnya sehingga membentuk satu *cluster* (Simamora, 2005).

Tujuan analisis cluster tidak dapat dipisahkan dengan pemilihan variabel yang akan digunakan untuk digolongkan objek ke dalam *cluster*. *Cluster* yang terbentuk merefleksikan struktur yang melekat pada data seperti yang didefinisikan oleh variabel-variabel. Pemilihan variabel harus sesuai dengan konsep yang umum digunakan dan harus rasional. Rasionalis ini didasarkan pada teori-teori eksplisit atau penelitian sebelumnya. Variabel-variabel yang dipilih hanyalah variabel yang dapat mencirikan objek yang akan dikelompokkan dan secara spesifik harus sesuai dengan tujuan analisis *cluster*.

Tahap-tahap pengclusteran data dengan metode hierarki yaitu:

a. Tentukan k sebagai jumlah cluster yang ingin dibentuk
b. Setiap data objek dianggap sebagai cluster. Kalau n = jumlah data dan c = jumlah cluster, berarti n = c
c. Menghitung jarak antar cluster
d. Menentukan dua cluster yang mempunyai jarak antar cluster paling kecil dan menggabungkannya (berarti n = c –1)
e. Jika n > k , maka kembali ke langkah 3.

Metode hierarki dilakukan dengan dua cara, yaitu metode penggabungan (*agglomerative*) dan metode pemisah (*devisive*) (Handoyo,2014).

a. Metode Penggabungan (*Agglomerative Method*)

Metode Penggabungan (*Agglomerative Method*) dilakukan dengan cara memandang setiap objek merupakan *cluster* yang berbeda. Kemudian setiap objek yang memiliki jarak terdekat akan digabungkan kedalam satu *cluster* dan objek yang memiliki jarak terdekat ketiga digabungkan kedalam *cluster* pertama atau bergabung dengan objek lain yang memiliki jarak terdekat yang sama sehinnga membentuk *cluster* baru. Tahap tersebut berlangsung sampai terbentuknya *cluster-cluster* yang terdiri dari keseluruhan objek. Metode Penggabungan *(agglomerative method*) terdiri dari lima metode, yaitu:

- Metode Pautan Tunggal (*Single Linkage Method*)

Single linkage method (jarak terdekat) atau tautan tunggal dapat dilakukan dengan mengelompokkan data berdasarkan jarak paling dekat (*nearest neighbour*) (Amponsah, et al, 2013). Dalam ukuran jarak terdekat mendefinisikan jarak antara dua cluster merupakan jarak terkecil antara *cluster* pertama dengan *cluster* kedua (Odilia, 2015). *Single linkage method* dapat dihitung dengan persamaan berikut (Rencher, 2002).

$$D(A,B) = min\{d(yi, yj), for\ yi\ in\ A\ and\ yj\ in\ B\}$$

- Metode Pautan Lengkap (*Complete Linkage Method*)

Complete linkage method (jarak jauh) dapat dilakukan dengan mengelompokkan data berdasarkkan jarak paling jauh. Ukuran ini sama dengan ukuran *single linkage*, namun yang membedakannya pada metode ini

memerlukan jarak terjauh antara *cluster* (Odilia, 2015). *Complete linkage method* dapat dihitung dengan persamaan berikut (Rencher, 2002).

$$D(A,B) = max\{d(yi, yj) \; for \; yi \; in \; A \; and \; yj \; in \; B\}.$$

- Metode Pautan Rata-Rata (*Average Linkage Method*)

  Average linkage method merupakan metode yang dilakukan dengan mengelompokkan data berdasarkan jarak rata-rata antar keseluruhan data. *Average linkage method* dapat dihitung dengan persamaan berikut (Rencher, 2002).

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i y_j)$$

- Metode Ward (*Ward's Method*)

  Ward's method adalah pengclusteran dengan memaksimumkan kesamaan dalam satu *cluster* serta memakai perhitungan yang lengkap. Pada setiap tahap, jarak antara dua *cluster* yang dapat dibentuk ialah *Sum of Square Error* (SSE) dalam dua *cluster* terkecil digabungkan (Rencher, 2002).

  Metode ward merupakan bagian dari metode hierarki yang membentuk pengclusteran sejumlah objek n ke dalam n, n-1, n-2, sampai seterusnya hingga membentuk satu *cluster*. *Sum of Square Error* (SSE) dapat lakukan jika setiap *cluster* beranggotakan lebih dari satu objek. *Sum of Square Error* (SSE) akan bernilai nol jika terdapat *cluster* yang beranggotakan satu objek saja. Metode *ward* dapat dihitung dengan persamaan berikut (Sarfia, 2016).

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})'(y_i - \bar{y})$$

Dimana:

y_i = nilai objek ke-i dengan i = 1, 2, 3, … n
$\bar{y}$ = rata-rata nilai objek dalam cluster
n = banyaknya cluster

Jika A, B dan AB merupakan *cluster*, maka jumlah kuadrat galat dalam cluster adalah:

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A)$$

$$SSE_B = \sum_{i=1}^{n_B} (y_i - \bar{y}_B)'(y_i - \bar{y}_B)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB})$$

Dimana $y_{AB} = (y_i y_i + y_i y_i)/(y_i + y_i)$ dan $y_i = y_i + y_i$ merupakan jumlah titik dalam *A, B,* dan *AB*. Karena, julah jarak setara jumlah cluster dalam kuadrat yang disimbolkan dengan $SSE_A, SSE_B \; dan \; SSE_{AB}$. Metode *ward* dapat bergabung dengan dua cluster A dan B yang dapat meminimalkan peningkatan *Sum of Square Error* (SSE).

- Metode Pusat (*Centroid Method*)

  Centroid method disebut juga metode titik pusat, dimana jarak antar *cluster* pada metode *centroid* merupakan jarak antar *centroid*. Jika terjadi pembentukan *cluster* baru maka akan terjadi perhitungan ulang (Rencher, 2002). Jarak antara dua kelompok didefinisikan sebagai berikut :

$$D(A, B) = d(\bar{y}_A, \bar{y}_B)$$

Dimana $\bar{y}_A$ dan $\bar{y}_B$ adalah vector rata-rata untuk pengamatan *cluster A* dan *B*, $\bar{y}_A$ dan $\bar{y}_B$ di definisikan: $\bar{y}_A = \sum_{i=1}^{n_A} y_1/n_A$. Dua cluster dengan jarak terkecil antara *centroid* bergabung pada setiap tahap.

Setelah dua *cluster A* dan *B* bergabung, pusat dari *cluster AB* dapat diberikan dengan rata-rata sebagai berikut:

$$\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}$$

2. Metode Pemecahan (*Devisive Method*)

Proses ini dimulai hanya dengan satu *cluster* saja, namun anggota cluster tersebut mencangkup semua objek penelitian. Kemudian objek memiliki perbedaan yang cukup besar akan pisahkan dan digabungkan ke dalam *cluster* yang sama. Proses ini berlangsung sampai terbentuknya sejumlah cluster yang diinginkan.

a. Metode Non Hierarki

Metode non hiearki disebut juga dengan metode k-means. Metode ini tidak sama halnya dengan metode hierarki, karena pada metode non hierarki dimulai dengan menentukan terlebih dahulu sejumlah *cluster* awal yang diinginkan, selanjutnya hasil dari objek pengamatan tersebut bergabung dan membentuk *cluster*. Pada metode non hieraki terdiri dari beberapa metode diantanya sebagai berikut(Gundono, 2011).

- Metode *Sequential Thereshold*

  Metode ini dimulai dengan memutuskan satu cluster dan menempatkan semua objek yang berada pada jarak tertentu ke dalam cluster. Jika semua objek telah digabungkan, selanjutnya menentukan cluster kedua dan

menempatkan semua objek yang berada pada jarak tertentu ke dalam cluster. Proses ini berlangsung seperti sebelumnya.

- Metode *Parallel Thereshold*
  Metode ini dimulai dengan memutuskan beberapa objek awal cluster secara bersamaan. Pada saat proses berlangsung, jarak yang memiliki kedekatan dapat ditentukan untuk memasukkan beberapa objek ke dalam cluster.
- Metode *Optimizing Partitionin*
  Metode ini serupa dengan dua metode sebelumnya. Namun yang membedakan pada metode ini memungkinkan untuk menempatkan kembali hasil objek-objek ke dalam cluster yang lebih dekat.

Tahap-tahap dalam pengclusteran data dengan metode non hierarki yaitu:

1) Menentukan banyaknya cluster yang ingin dibentuk, misalnya sebanyak cluster.
2) Tentukan pusat cluster (bisa ditentukan secara sembarang). Hal ini termasuk salah satu kelemahan dalam metode non hierarki.
3) Mengalokasikan objek ke cluster yang terdekat dengan pusat cluster.
4) Pusat cluster dihitung kembali, yang merupakan rata-rata dari individu didalam kelompok itu sendiri.
5) Alokasikan kembali individu. Tahap berlangsung hingga tidak ada lagi objek yang berpindah cluster (Rencher, 2002).

2.2.3 Menentukan Banyak Cluster

Isi pokok pada analisis cluster yaitu memutuskan seberapa banyaknya cluster, namun dalam menentukan jumlah cluster tidak ada aturan secara baku. Dalam hal ini ada beberapa petunjuk yang bisa digunakan diantaranya sebagai berikut (Supranto, 2004):

a. Pertimbangan teoretis, konseptual, praktis, mungkin bisa diusulkan/disarankan untuk menentukan berapa banyaknya cluster yang sebenarnya. Sebagai contoh, kalau tujuan pengclusteran untuk mengenali/mengidentifikasi segmen pasar, manajemen mungkin menghendaki cluster dalam jumlah tertentu (katakan 3, 4, atau 5 cluster).
b. Di dalam pengclusteran hierarki, jarak dimana cluster digabung bisa dipergunakan sebagai kriteria.
c. Di dalam pengclusteran non hierarki, rasio jumlah varian dalam cluster dengan jumlah varian antar cluster dapat diplotkan melawan banyaknya cluster.
d. Besarnya relatif cluster seharusnya berguna/bermanfaat.

2.2.3.1 Metriks Evaluasi

Penjelasan singkat untuk beberapa metrik evaluasi umum dalam konteks *clustering*:

a. *Silhouette Score*
   Mengukur sejauh mana objek-objek dalam klaster sesuai dengan klasternya sendiri dibandingkan dengan klaster tetangga terdekatnya. Rentang nilai: -1 hingga 1. Nilai positif menunjukkan kualitas klastering yang baik.
b. *Davies-Bouldin Index*

Mengukur seberapa baik setiap klaster terisolasi dari klaster lainnya. Semakin rendah nilai, semakin baik klasteringnya. Rentang nilai: 0 hingga ∞. Nilai 0 menunjukkan klastering yang sempurna.

c. *Calinski-Harabasz Index (Variance Ratio Criterion)*
Menilai kekompakan intra-klaster dan dispersi inter-klaster. Semakin tinggi nilainya, semakin baik kualitas klasteringnya. Rentang nilai: Semakin tinggi, semakin baik.

d. *Adjusted Rand Index (*ARI*):*
Mengukur seberapa baik klastering sesuai dengan klaster referensi (ground truth), memperhitungkan kemungkinan kebetulan. Rentang nilai: -1 hingga 1. Nilai positif menunjukkan kesesuaian yang baik.

2.2.4 Principal Component Analysis (PCA)

Menurut (Supranto, 2004), PCA merupakan suatu teknik mereduksi data multivariat (banyak data) yang mencari untuk mengubah (mentransformasi) suatu matriks data awal/asli menjadi satu himpunan kombinasi linier yang lebih sedikit, akan tetapi menyerap sebagian besar jumlah varians dari data awal. Tujuan utama dari PCA adalah menjelaskan sebanyak mungkin jumlah varians data asli dengan sedikit mungkin komponen utama yang disebut faktor. Banyaknya faktor (komponen) yang dapat diekstrak dari data awal/asli adalah sebanyak variabel yang ada.

Selain itu, (Yamin, 2011) juga menjelaskan PCA pada dasarnya teknik statistik yang bertujuan untuk menyederhanakan variabel yang diamati dengan cara mereduksi dimensinya (disebut juga sebagai teknik pereduksian data). Prinsip utama dalam PCA adalah terdapatnya korelasi di antara variabel. Apabila hal ini terjadi, maka ada estimasi peneliti bahwa sesungguhnya beberapa variabel tersebut dapat.

2.2.5 Interpretasi Suatu Cluster

Interpretasi suatu cluster merupakan nilai yang digunakan untuk mengetahui karakteristik masing-masing cluster agar dapat menjelaskan bagaimana perbedaan yang terjadi dari setiap cluster secara relevan. Ukuran yang bisa digunakan untuk proses interpretasi ini adalah menghitung centroid atau rata-rata variabel yang merupakan karakteristik masing-masing objek pada setiap cluster. Rumus yang digunakan adalah sebagai berikut.

$$v = \frac{\sum_{i=i}^{n} yi}{n}$$

Keterangan:
v : Nilai centroid atau nilai rata-rata
yi : Objek ke-i
n : Banyak objek

BAB III

METODOLOGI PENELITIAN

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data sekunder. Dimana data di dapat tidak secara langsung dari objek penelitian. Peneliti mendapatkan data yang sudah jadi yang dikumpulkan oleh pihak lain dengan berbagai cara atau metode baik secara komersial maupun non komersial. Data sekunder ini meliputi pengeluaran seminggu kelompok kacang per kabupaten/kota di Indonesia tahun 2021 sampai 2022, yang peroleh dari Badan Pusat Statistika (BPS) dengan subjek Konsumsi dan Pengeluaran.

3.2 Analisis

Metode analisis pada penelitian ini yaitu *analytics descriptive* untuk melihat gambaran umum mengenai data pengeluaran seminggu kelompok kacang per kabupaten/kota di Indonesia tahun 2021 sampai 2022, kemudian dilanjutkan dengan analisis mengetahui pengelompokan wilayah-wilayah berdasarkan rata-rata pengeluaran perkapita seminggu pada kacang-kacangan dan kemudian memberikan pemahaman tentang pola konsumsi atau struktur ekonomi yang mungkin terkait dengan setiap kelompok. Pada tahapan ini juga gambaran tentang penelitian yang akan dilakukan telah diketahui. Terdapat dua proses analisis yang dilakukan, yaitu analisis terhadap data dan analisis terhadap tahap KDD.

Analisis data dilakukan terhadap data pengeluaran disetiap kabupaten/kota. pengeluaran yang diambil mulai dari data periode 2021 sampai periode 2022. Data tersebut akan melalui tahapan KDD yaitu *data cleaning* dan *data selection* . Tahapan tersebut dilakuan bertujuan untuk mendapatkan data yang valid untuk dijadikan dan digunakan dalam penelitian ini. Setelah tahapan tersebut dilakukan, selanjutnya akan masuk pada tahap *data mining*. Pada tahap ini akan menggunakan *hierarchical clustering* dengan metode *average linkage.*

3.3 Langkah Analisis Data

Adapun langkah-langkah analisis data untuk melakukan penelitian tentang pengelompokan pengeluaran untuk kacang kabupaten/kota tahun 2021-2022 antara lain:

1. Mengelompokan pengeluaran untuk kacang dengan langkah-langkah analisis sebagai berikut:
   a. Melakukan analisis lebih lanjut yang dapat menggambarkan kelompok
   b. Memperoleh hasil dengan algoritma *hierarchical clustering* dengan metode *average linkage*
2. Interpretasi
   Pada tahap menginterpretasikan terdapat pengujian terhadap masing-masing cluster yang telah terbentuk. Kemudian proses ini digunakan dalam mendeskripsikan karakteristik dari setiap hasil cluster pada profil tertentu secara tepat untuk mengetahui hasil dari cluster yang terbentuk. Pada proses menginterpretasikan menggunakan rata-rata (centroid) di semua anggota dalam cluster.

3. Menarik Kesimpulan
   Tahap ini berisikan tentang kesimpulan penelitian ini dan hasil yang didapatkan. Tahap ini juga berisikan hal yang disimpulkan dan disarankan penulis bagi pembaca untuk melakukan pengembangan terhadap penelitian ini kedepannya.

BAB IV

ANALISA DAN PEMBAHASAN

Secara garis besar model yang dirancang pada penelitian ini terdiri dari dua bagian utama, yaitu tahapan *preprocessing data* dan tahapan *data mining* (*hierarchical clustering*). Langkah pertama yang harus dilakukan sebelum pembuatan model adalah mengumpulkan data penelitian. Data yang telah dikumpulkan akan dilakukan tahapan *preprocessing data*. Data hasil *preprocessing* kemudian akan diolah dengan menggunakan metode algoritma *hierarchical clustering*, sehingga menghasilkan model berupa plotting. Dalam upaya untuk mengelola dan menganalisis data dengan skala besar, pilihan untuk menggunakan *Python* dengan *PySpark* menjadi sangat relevan.

4.1 Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data pengeluaran seminggu kelompok kacang per kabupaten/kota di Indonesia tahun 2021 sampai 2022. Dataset tersebut berjumlah 514 baris dan 17 kolom. Data ini diambil dari website resmi Badan Pusat Statistika (BPS). Data memiliki atribut Kabupaten/Kota, 2021_Kacang-kacangan, 2022_Kacang-kacangan, 2021_Kacang tanah tanpa kulit, 2022_Kacang tanah tanpa kulit, 2021_Kacang kedele, 2022_Kacang kedele, 2021_Kacang lainnya, 2022_Kacang lainnya, 2021_tahu, 2022_tahu, 2021_tempe, 2022_tempe, 2021_oncom, 2022_oncom, 2021_Hasil lain dari kacang-kacangan, 2022_Hasil lain dari kacang-kacangan. Atribut-atribut tersebut akan dikelola. Berikut adalah Tabel 1 data pengeluaran kelompok kacang.

Kabupaten/Kota	Rata-rata Pengeluaran Perkapita Seminggu Menurut Kelompok Kacang-Kacangan Per Kabupaten/kota (Rupiah/Kapita/Minggu)															
	Kacang-kacangan		Kacang tanah tanpa kulit		Kacang kedele		Kacang lainnya		Tahu		Tempe		Oncom		Hasil lain dari kacang-kacangan	
	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022
Simeulue	475,28	582,48	44,9	18,95	-	-	33,89	56,8	79,96	83,12	316,52	421,32	-	2,28	-	-
Aceh Singkil	2033,93	2189,15	103,43	99,07	2,28	18,45	38,45	14,07	947,5	1032,3	928,08	1020,29	0,73	4,98	13,45	-
Aceh Selatan	1324,44	1016,84	235,14	70,04	-	-	45,25	10,67	302,39	293,1	730,39	635,83	-	5,6	11,27	1,59
Aceh Tenggara	1812,45	1864,99	87,29	30,43	2	20,39	21,84	3,72	846,88	908,09	822,4	861,9	16,94	36,99	15,11	3,47
Aceh Timur	1329,52	1482,39	34,91	85,69	-	16,5	39,6	52,32	460,16	504,28	765,32	800,32	13,48	19,94	16,05	3,34
Aceh Tengah	2953,51	2982,12	87,83	63,48	21,48	-	87,33	16,4	1432,48	1500,87	1324,37	1377,15	-	24,23	-	-
Aceh Barat	2187,72	2194,79	279,74	148,66	28,66	20,18	85,85	104,59	605,68	648,31	1175,07	1264,72	5,4	4,64	7,32	3,68
Aceh Besar	1776,45	1774,67	103,54	130,36	14,49	11,63	39,15	108,75	632,05	612,08	986,4	904,44	0,83	7,42	-	-
Pidie	1470,29	1471,89	241,77	80,72	2,06	12,65	122,6	75,54	345,14	439,71	748,97	855,85	9,29	0,95	0,46	6,47
Bireuen	2008,42	1708,89	154,82	62,55	19,51	28,77	224,49	71,68	580,03	530,34	1005,62	991,58	6,41	11,11	17,54	12,86
Aceh Utara	1228,4	1556,62	65,29	110,09	16,09	4,58	67,63	31,05	376,62	540,04	672,86	852,59	15,11	17	14,8	1,26
Aceh Barat Daya	1094,49	950,95	111,5	50,52	3,18	9,5	5,5	14,29	322,72	294,05	648,59	580,55	3	0,9	-	1,14
Gayo Lues	2002,71	1973,6	213,11	238,91	12,69	7,75	79,61	11,21	652,69	690,29	1025,81	1000,56	13,5	18,39	5,29	6,48
Aceh Tamiang	1955,06	1990,74	32,19	39,95	6,7	4,12	21,19	6,92	824,51	778,42	1061,15	1147,27	8,86	13,66	0,46	0,4
Nagan Raya	2614,43	2965,64	280,12	273,09	28,62	2,95	104,03	94,99	848,85	1038,45	1331,13	1515,27	17,7	35,6	3,98	5,29
Aceh Jaya	1694,25	1655,78	238,53	160,17	-	3,89	84,39	71,34	405,4	441,67	962,52	972,89	0,61	0,61	2,8	5,21
Bener Meriah	2496,53	2847,19	78,02	111,82	6,78	10,8	8,15	13,03	1223,05	1142,48	1168,26	1524,94	12,27	44,13	-	-
Pidie Jaya	1838,26	1742,77	226,58	52,29	-	12,91	80,57	21,55	542,01	544,99	968,84	1088,06	15,76	22,97	4,51	-
Kota Banda Aceh	2192,73	2720,31	163,1	176,49	3,66	34,6	64,33	59,5	719,8	891,01	1224,83	1535,2	4,29	0,56	12,72	22,96
Kota Sabang	2113,01	2416,49	49,65	60,41	5,36	22,84	64,18	157,25	769,2	911,13	1214,59	1239,48	9,31	2,95	0,72	22,44
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Kota Jayapura	3134,49	4085,36	55,16	101,56	-	-	14,21	126,04	1924,05	2123,06	1141,07	1729,02	-	5,69	-	-

Tabel 1. Data Pengeluaran Kelompok Kacang

Data yang sudah dikumpulkan selanjutnya akan dilakukan tahap *preprocessing data* dan tahap tahapan *data mining* yang ada didalam tahapan KDD.

4.2 Preprocessing Data

4.2.1 Data Cleaning

Pada umumnya, data yang diperoleh, baik dari database suatu perusahaan maupun eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau hanya sekedar salah ketik. Data yang tidak relevan itu juga lebih baik dibuang karena keberadaanya bisa mengurangi mutu atau akurasi dari hasil data mining nantinya. Pembersihan data juga akan mempengaruhi performasi dari system data mining karena data yang akan ditangani akan berkurang jumlah dan kompleksitasnya. Berdasarkan pengecekan pada data kelulusan, tidak ditemukan data yang hilang dan tidak valid, sehingga tahapan cleaning data tidak dilakukan.

Handling data yang dilakukan oleh peneliti adalah mengisi *missing value* (nilai yang hilang). Dimana nilai yang hilang pada suatu dataset diganti atau diisi dengan nilai tertentu. Ini penting karena nilai yang hilang dapat mempengaruhi analisis data dan kinerja model. Metode untuk mengisi *missing value* adalah mengisi nilai yang hilang dengan rata-rata (*mean*). Metode ini sering digunakan jika data terdistribusi normal atau memiliki nilai yang dominan.

4.2.2 Data Selection

Proses pada tahap ini adalah menseleksi atribut data yang akan dipergunakan. Hanya atribut yang dibutuhkan dan relevan saja. Berdasarkan analisis maka total atribut yang digunakan dalam penelitian ini adalah lima belas atribut yaitu Kabupaten/Kota, 2021_Kacang-kacangan, 2022_Kacang-kacangan, 2021_Kacang tanah tanpa kulit, 2022_Kacang tanah tanpa kulit, 2021_Kacang kedele, 2022_Kacang kedele, 2021_Kacang lainnya, 2022_Kacang lainnya, 2021_tahu, 2022_tahu, 2021_tempe, 2022_tempe, 2021_oncom, 2022_oncom. Data selection dapat dilihat pada Tabel 2 berikut:

Kabupaten/Kota	2021_Kacang-kacangan	2022_Kacang-kacangan	2021_Kacang tanah tanpa kulit	2022_Kacang tanah tanpa kulit	2021_Kacang kedele	2022_Kacang kedele	2021_Kacang lainnya	2022_Kacang lainnya	2021_tahu	2022_tahu	2021_tempe	2022_tempe	2021_oncom	2022_oncom
Simeulue	475,28	582,48	44,9	18,95	14,02	15,37	33,89	56,8	79,96	83,12	316,52	421,32	21,3	2,28
Aceh Singkil	2033,93	2189,15	103,43	99,07	2,28	18,45	38,45	14,07	947,5	1032,3	928,08	1020,29	0,73	4,98
Aceh Selatan	1324,44	1016,84	235,14	70,04	14,02	15,37	45,25	10,67	302,39	293,1	730,39	635,83	21,3	5,6
Aceh Tenggara	1812,45	1864,99	87,29	30,43	2	20,39	21,84	3,72	846,88	908,09	822,4	861,9	16,94	36,99
Aceh Timur	1329,52	1482,39	34,91	85,69	14,02	16,5	39,6	52,32	460,16	504,28	765,32	800,32	13,48	19,94
Aceh Tengah	2953,51	2982,12	87,83	63,48	21,48	15,37	87,33	16,4	1432,48	1500,87	1324,37	1377,15	21,3	24,23
Aceh Barat	2187,72	2194,79	279,74	148,66	28,66	20,18	85,85	104,59	605,68	648,31	1175,07	1264,72	5,4	4,64
Aceh Besar	1776,45	1774,67	103,54	130,36	14,49	11,63	39,15	108,75	632,05	612,08	986,4	904,44	0,83	7,42
Pidie	1470,29	1471,89	241,77	80,72	2,06	12,65	122,6	75,54	345,14	439,71	748,97	855,85	9,29	0,95
Bireuen	2008,42	1708,89	154,82	62,55	19,51	28,77	224,49	71,68	580,03	530,34	1005,62	991,58	6,41	11,11
Aceh Utara	1228,4	1556,62	65,29	110,09	16,09	4,58	67,63	31,05	376,62	540,04	672,86	852,59	15,11	17
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Kota Jayapura	3134,49	4085,36	55,16	101,56	14,02	15,37	14,21	126,04	1924,05	2123,06	1141,07	859,96	21,3	3,79

Tabel 2. Data Hasil Seleksi

4.3 Proses Cluster Menggunakan Hierarchical Clustering Metode Average Linkage

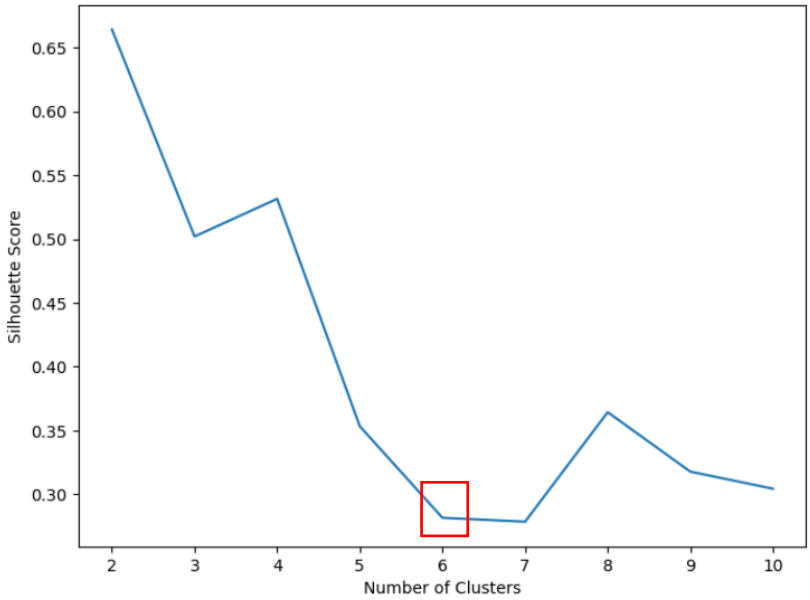
Berikut adalah langkah-langkah pada *hierarchical clustering*:

1. Merangkai 14 variabel kolom menjadi satu kolom vector fitur ‘features’ menggunakan ‘VectorAssembler’ dari Pyspark.

features
[475.28,582.48,44...]
[2033.93,2189.15,...]
[1324.44,1016.84,...]
[1812.45,1864.99,...]
[1329.52,1482.39,...]
[2953.51,2982.12,...]
[2187.72,2194.79,...]
[1776.45,1774.67,...]
[1470.29,1471.89,...]
[.....,.....,...]
[1228.4,1556.62,6...]

Tabel 3. Kolom Vector

2. Memutuskan seberapa banyak *cluster* yang dipilih. Proses pengclusteran pada metode agglomeratif (*agglomerative method*) dengan metode *average* dilakukan dengan mengelompokan data berdasarkan jarak rata-rata antar keseluruhan data. Tahap selanjutnya dalam proses *cluster* ialah tentukan k sebagai jumlah cluster yang ingin dibentuk. Untuk hal ini dapat dibantu dengan metriks evaluasi yaitu *silhouette score*. Nilai *silhouette score* dipilih saat garis siku terbentuk. Menetapkan k=6.



Gambar 1. Silhouette Score

3. Setelah itu, membangun model *hierarchical clustering* dengan menyesuaikan data cluster yang didapat. Kumpulan data pada dataset disesuaikan dengan algoritma *hierarchical clustering* dan buat kolom baru yang disebut 'prediction'.

features	prediction
[475.28,582.48,44...	0
[2033.93,2189.15,...	1
[1324.44,1016.84,...	0
[1812.45,1864.99,...	1
[1329.52,1482.39,...	0
[2953.51,2982.12,...	3
[2187.72,2194.79,...	1
[1776.45,1774.67,...	1
[1470.29,1471.89,...	0
[2008.42,1708.89,...	1
[1228.4,1556.62,6...	0
[1094.49,950.95,1...	0
[……,……,...	…
[3134.49,4085.36,...	4

Tabel 4. Kolom Features dan Prediction

4.4 Analisis Hierarchical Clustering

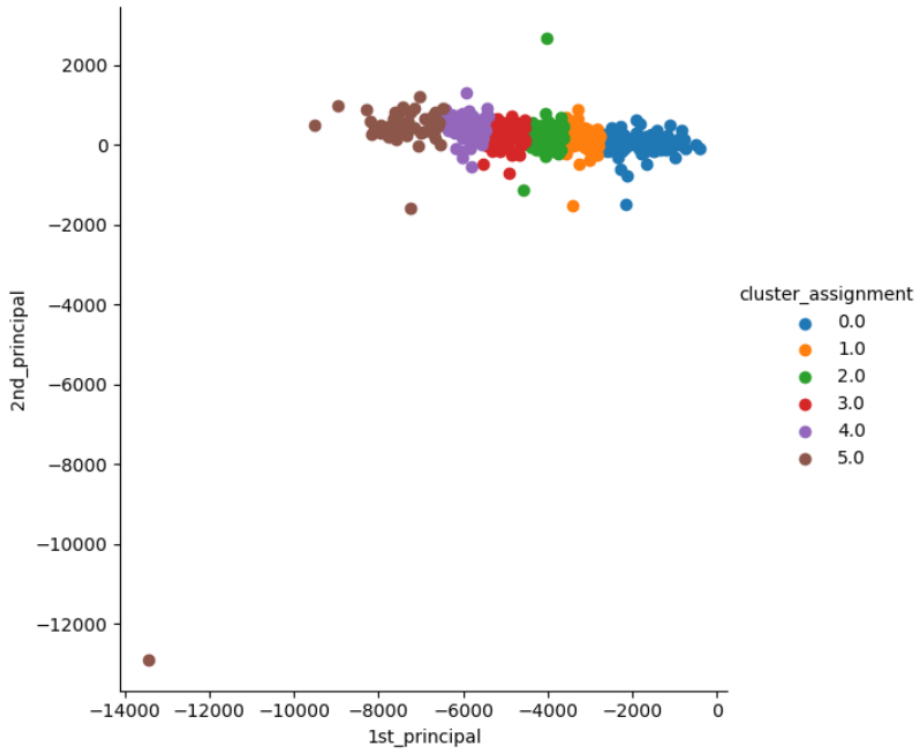
Analisis *hierarchical clustering* merupakan suatu metode yang tidak membutuhkan suatu asumsi yang dibuat dalam jumlah kelompok atau struktur kelompok. Analisis *hierarchical clustering* sendiri merupakan suatu metode pengelompokan yang jumlah kelompok yang akan dibuat belum diketahui. Dalam analisis ini, data dikelompokkan berdasarkan kemiripan atau kedekatan antara observasi-observasi yang ada dalam dataset. Proses pengclusteran dapat dilakukan dengan vector fitur.

Principal Component Analysis (PCA) dilakukan membuat plot cluster ke dalam diagram 2D untuk visibilitas yang baik. Karena saat ini, terdapat 14 kolom/fitur/dimensi diperoleh dari 2021_Kacang-kacangan, 2022_Kacang-kacangan, 2021_Kacang tanah tanpa kulit, 2022_Kacang tanah tanpa kulit, 2021_Kacang kedele, 2022_Kacang kedele, 2021_Kacang lainnya, 2022_Kacang lainnya, 2021_tahu, 2022_tahu, 2021_tempe, 2022_tempe, 2021_oncom, 2022_oncom. Memvisualisasikan data 14 dimensi ini menjadi 2 dimensi, oleh karena itu diperlukan PCA. Berikut hasilnya:

	pca
0	[-883.9019335532776, -4.090849141278189]
1	[-3572.4030427258976, 156.3869428602995]
2	[-1908.8447645937076, 287.6619978510805]
3	[-3107.700932627474, 218.78862156085737]
4	[-2366.4696155577267, 69.30341559097445]
..	...
509	[-490.2099665697026, 17.103862829649998]
510	[-2119.143475660604, -762.5228317851236]
511	[-1115.6895673907425, -73.13545285946715]
512	[-1428.2929186734902, 351.6555452141348]
513	[-6187.315758381206, -99.18316611543626]

Tabel 4. Kolom Data PCA

Berdasarkan hasil PCA tersebut dapat membuat plotting.



Gambar 2. Plotting PCA with Prediction

Dari Gambar 2 diketahui bahwa terbentuk enam cluster sebagai berikut.

Cluster	Anggota
Cluster 0	'Simeulue', 'Aceh Selatan', 'Aceh Timur', 'Pidie', 'Aceh Utara', 'Aceh Barat Daya', 'Aceh Jaya', 'Nias', 'Tapanuli Tengah', 'Nias Selatan', 'Batu Bara', 'Nias Utara', 'Nias Barat', 'Kota Sibolga', 'Kota Tanjung Balai', 'Kota Gunungsitoli', 'Kepulauan Mentawai', 'Bima', 'Sumba Barat', 'Sumba Timur', 'Alor', 'Lembata', 'Flores Timur', 'Sikka', 'Rote Ndao', 'Manggarai Barat', 'Sumba Tengah', 'Sumba Barat Daya', 'Nagekeo', 'Manggarai Timur', 'Kapuas Hulu', 'Kepulauan Sangihe', 'Siau Tagulandang Biaro', 'Minahasa Tenggara', 'Bolaang Mongondow Selatan', 'Banggai Kepulauan', 'Buol', 'Tojo Una-Una', 'Banggai Laut', 'Kepulauan Selayar', 'Bantaeng', 'Takalar', 'Sinjai', 'Bone', 'Soppeng', 'Wajo', 'Pinrang', 'Luwu', 'Tana Toraja', 'Buton', 'Muna', 'Bombana', 'Wakatobi', 'Kolaka Utara', 'Buton Utara', 'Konawe Kepulauan', 'Muna Barat', 'Buton Tengah', 'Buton Selatan', 'Kota Baubau', 'Pohuwato', 'Bone Bolango', 'Gorontalo Utara', 'Majene', 'Polewali Mandar', 'Mamasa', 'Mamuju', 'Maluku Tenggara Barat', 'Maluku Tenggara', 'Kepulauan Aru', 'Seram Bagian Barat', 'Seram Bagian Timur', 'Maluku Barat Daya', 'Buru Selatan', 'Kota Tual', 'Halmahera Barat', 'Halmahera Tengah', 'Kepulauan Sula', 'Halmahera

	Selatan', 'Halmahera Utara', 'Pulau Morotai', 'Pulau Taliabu', 'Kota Tidore Kepulauan', 'Kaimana', 'Sorong Selatan', 'Raja Ampat', 'Tambrauw', 'Pegunungan Arfak', 'Mappi', 'Asmat', 'Yahukimo', 'Pegunungan Bintang', 'Sarmi', 'Mamberamo Raya', 'Lanny Jaya', 'Mamberamo Tengah', 'Puncak', 'Dogiyai', 'Intan Jaya', 'Deiyai'
Cluster 1	'Aceh Singkil', 'Aceh Tenggara', 'Aceh Barat', 'Aceh Besar', 'Bireuen', 'Gayo Lues', 'Aceh Tamiang', 'Pidie Jaya', 'Kota Lhokseumawe', 'Mandailing Natal', 'Labuhan Batu', 'Humbang Hasundutan', 'Samosir', 'Serdang Bedagai', 'Labuhan Batu Utara', 'Kota Pematang Siantar', 'Kota Tebing Tinggi', 'Pesisir Selatan', 'Sijunjung', 'Tanah Datar', 'Padang Pariaman', 'Agam', 'Lima Puluh Kota', 'Pasaman', 'Pasaman Barat', 'Kota Bukittinggi', 'Kota Payakumbuh', 'Kota Pariaman', 'Indragiri Hilir', 'Merangin', 'Bungo', 'Kota Sungai Penuh', 'Ogan Komering Ulu Selatan', 'Ogan Ilir', 'Empat Lawang', 'Penukal Abab Lematang Ilir', 'Musi Rawas Utara', 'Kota Prabumulih', 'Kota Lubuklinggau', 'Kaur', 'Bengkulu Tengah', 'Bangka Selatan', 'Natuna', 'Lingga', 'Kepulauan Anambas', 'Karangasem', 'Dompu', 'Kupang', 'Ende', 'Ngada', 'Sambas', 'Mempawah', 'Sintang', 'Sekadau', 'Kubu Raya', 'Kapuas', 'Gunung Mas', 'Banjar', 'Barito Kuala', 'Hulu Sungai Selatan', 'Hulu Sungai Utara', 'Bolaang Mongondow', 'Minahasa', 'Kepulauan Talaud', 'Minahasa Selatan', 'Bolaang Mongondow Utara', 'Bolaang Mongondow Timur', 'Kota Kotamobagu', 'Morowali', 'Donggala', 'Toli-Toli', 'Parigi Moutong', 'Bulukumba', 'Jeneponto', 'Gowa', 'Maros', 'Pangkajene dan Kepulauan', 'Barru', 'Sidenreng Rappang', 'Luwu Utara', 'Luwu Timur', 'Toraja Utara', 'Kota Parepare', 'Kota Palopo', 'Konawe', 'Kolaka', 'Konawe Utara', 'Boalemo', 'Gorontalo', 'Kota Gorontalo', 'Mamuju Tengah', 'Maluku Tengah', 'Buru', 'Halmahera Timur', 'Kota Ternate', 'Teluk Wondama', 'Maybrat', 'Kepulauan Yapen', 'Boven Digoel', 'Tolikara''Waropen'
Cluster 2	'Kota Banda Aceh', 'Kota Sabang', 'Kota Langsa', 'Kota Subulussalam', 'Tapanuli Selatan', 'Tapanuli Utara', 'Toba Samosir', 'Asahan', 'Simalungun', 'Dairi', 'Deli Serdang', 'Langkat', 'Pakpak Bharat', 'Padang Lawas Utara', 'Padang Lawas', 'Labuhan Batu Selatan', 'Kota Medan', 'Kota Binjai', 'Kota Padangsidimpuan', 'Solok', 'Solok Selatan', 'Dharmasraya', 'Kota Padang', 'Kota Solok', 'Kota Padang Panjang', 'Kuantan Singingi', 'Indragiri Hulu', 'Siak', 'Kampar', 'Rokan Hulu', 'Bengkalis', 'Rokan Hilir', 'Kepulauan Meranti', 'Kota Pekanbaru', 'Kerinci', 'Sarolangun', 'Batang Hari', 'Tanjung Jabung Barat', 'Tebo', 'Ogan Komering Ulu', 'Ogan Komering Ilir', 'Muara Enim', 'Lahat', 'Musi Rawas', 'Kota Pagar Alam', 'Bengkulu Selatan', 'Seluma', 'Mukomuko', 'Lebong', 'Kepahiang', 'Kota Bengkulu', 'Tulang Bawang Barat', 'Pesisir Barat', 'Bangka', 'Belitung', 'Bangka Barat', 'Bangka Tengah', 'Belitung Timur', 'Karimun', 'Kep. Seribu', 'Cianjur', 'Kota Tasikmalaya', 'Pati', 'Pamekasan', 'Pandeglang', 'Buleleng', 'Sumbawa', 'Kota Bima', 'Timor Tengah Selatan', 'Belu', 'Manggarai', 'Malaka', 'Bengkayang', 'Landak', 'Sanggau', 'Ketapang', 'Melawi', 'Kayong Utara', 'Barito Selatan', 'Barito Utara', 'Katingan', 'Barito Timur', 'Murung Raya', 'Hulu Sungai Tengah', 'Balangan', 'Kota Banjarmasin', 'Malinau', 'Nunukan', 'Minahasa Utara', 'Kota Manado', 'Kota Bitung', 'Kota Tomohon', 'Banggai', 'Poso', 'Sigi', 'Morowali Utara', 'Konawe Selatan', 'Kolaka Timur', 'Kota Kendari', 'Mamuju Utara', 'Kota Ambon', 'Fakfak', 'Kota Sorong', 'Biak Numfor', 'Paniai', 'Supiori', 'Yalimo'

Cluster 3	'Aceh Tengah', 'Nagan Raya', 'Bener Meriah', 'Karo', 'Kota Sawah Lunto', 'Pelalawan', 'Kota Dumai', 'Muaro Jambi', 'Tanjung Jabung Timur', 'Kota J ambi', 'Musi Banyuasin', 'Banyu Asin', 'Ogan Komering Ulu Timur', 'Kota Palembang', 'Rejang Lebong', 'Bengkulu Utara', 'Lampung Barat', 'Tangga mus', 'Lampung Utara', 'Way Kanan', 'Tulangbawang', 'Pesawaran', 'Kota Pangkal Pinang', 'Bintan', 'Kota Batam', 'Kota Tanjung Pinang', 'Bogor', 'S ukabumi', 'Bandung', 'Garut', 'Tasikmalaya', 'Majalengka', 'Sumedang', 'Su bang', 'Karawang', 'Bandung Barat', 'Pangandaran', 'Kota Sukabumi', 'Kot a Cirebon', 'Kota Banjar', 'Cilacap', 'Banyumas', 'Temanggung', 'Kota Pek alongan', 'Kota Tegal', 'Kulon Progo', 'Kota Yogyakarta', 'Trenggalek', 'Bl itar', 'Lamongan', 'Bangkalan', 'Tangerang', 'Serang', 'Jembrana', 'Tabanan' , 'Klungkung', 'Lombok Tengah', 'Sumbawa Barat', 'Timor Tengah Utara', 'Kota Pontianak', 'Kota Singkawang', 'Tanah Laut', 'Kota Baru', 'Tapin', 'T abalong', 'Kutai Barat', 'Mahakam Ulu', 'Kota Palu', 'Kota Makasar', 'Soro ng', 'Manokwari Selatan', 'Jayawijaya', 'Mimika'
Cluster 4	'Lampung Timur', 'Lampung Tengah', 'Pringsewu', 'Mesuji', 'Kota Metro', 'Kota Jakarta Timur', 'Kota Jakarta Pusat', 'Kota Jakarta Barat', 'Kota Jaka rta Utara', 'Ciamis', 'Cirebon', 'Purwakarta', 'Bekasi', 'Kota Bogor', 'Purbal ingga', 'Banjarnegara', 'Purworejo', 'Wonosobo', 'Magelang', 'Klaten', 'Suk oharjo', 'Grobogan', 'Blora', 'Rembang', 'Kudus', 'Jepara', 'Demak', 'Kendal ', 'Batang', 'Pekalongan', 'Pemalang', 'Brebes', 'Bantul', 'Sleman', 'Pacitan', 'Ponorogo', 'Tulungagung', 'Malang', 'Banyuwangi', 'Ngawi', 'Bojonegoro' , 'Tuban', 'Gresik', 'Sampang', 'Kota Blitar', 'Lebak', 'Kota Cilegon', 'Badu ng', 'Gianyar', 'Bangli', 'Lombok Barat', 'Lombok Utara', 'Sabu Raijua', 'K ota Kupang', 'Kotawaringin Timur', 'Sukamara', 'Seruyan', 'Pulang Pisau', 'Kota Palangka Raya', 'Tanah Bumbu', 'Kota Banjar Baru', 'Paser', 'Kutai Kartanegara', 'Penajam Paser Utara', 'Kota Samarinda', 'Kota Bontang', 'B ulungan', 'Tana Tidung', 'Kota Tarakan', 'Enrekang', 'Teluk Bintuni', 'Man okwari', 'Jayapura', 'Keerom', 'Nduga', 'Kota Jayapura'
Cluster 5	'Lampung Selatan', 'Kota Bandar Lampung', 'Kota Jakarta Selatan', 'Kunin gan', 'Indramayu', 'Kota Bandung', 'Kota Bekasi', 'Kota Depok', 'Kota Cim ahi', 'Kebumen', 'Boyolali', 'Wonogiri', 'Karanganyar', 'Sragen', 'Semarang ', 'Tegal', 'Kota Magelang', 'Kota Surakarta', 'Kota Salatiga', 'Kota Semara ng', 'Gunung Kidul', 'Kediri', 'Lumajang', 'Jember', 'Bondowoso', 'Situbon do', 'Probolinggo', 'Pasuruan', 'Sidoarjo', 'Mojokerto', 'Jombang', 'Nganjuk' , 'Madiun', 'Magetan', 'Sumenep', 'Kota Kediri', 'Kota Malang', 'Kota Prob olinggo', 'Kota Pasuruan', 'Kota Mojokerto', 'Kota Madiun', 'Kota Surabay a', 'Kota Batu', 'Kota Tangerang', 'Kota Serang', 'Kota Tangerang Selatan', 'Kota Denpasar', 'Lombok Timur', 'Kota Mataram', 'Kotawaringin Barat', ' Lamandau', 'Kutai Timur', 'Berau', 'Kota Balikpapan', 'Merauke', 'Nabire', 'Puncak Jaya'

Tabel 5. Anggota Cluster Average Linkage

Pada Tabel 5 dapat diketahui masing-masing dari anggota cluster. Pada cluster 2 merupakan cluster dengan jumlah berbesar yaitu 107 objek. Urutan cluster terbesar kedua yaitu cluster 1 dengan 101 objek. Lalu, terbesar ketiga yaitu cluster 0 dengan 100 objek. Urutan keempat

cluster 4 dengan 76 objek. Setelah, itu dua urutan cluster terkecil yaitu cluster 3 dengan 73 objek dan cluster 5 dengan 57 objek Kabupaten/Kota.

4.5 Menginterpretasi Suatu Cluster

Interpretasi cluster yang terbentuk dengan menghitung rataan variabel pada setiap objek yang terdapat di setiap cluster. Hasil ini dapat memberikan gambaran mengenai karakteristik masing-masing cluster yang terbentuk dari proses pengelompokan dengan metode *average linkage*. Berikut ini adalah hasil rataan untuk interpretasi cluster *average linkage*.

Variabel	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
2021_Kacang-kacangan	1154.2	1922.81	2358.23	2859.36	3481.89	4226.22
2022_Kacang-kacangan	1171.66	1970.62	2452.44	3066.23	3515.11	4630.88
2021_Kacang tanah tanpa kulit	171.43	166.48	125.6	125.49	141.62	171.77
2022_Kacang tanah tanpa kulit	148.94	140.18	139.46	128.22	126.11	316.22
2021_Kacang kedele	16.56	10.93	10.98	14.1	14.14	20.5
2022_Kacang kedele	15.52	14.25	12.18	15.08	15.86	22.82
2021_Kacang lainnya	78.43	56.16	75.39	67.9	108.65	72.92
2022_Kacang lainnya	94.37	42.52	71.43	55.44	71.89	60.83
2021_tahu	491.6	875.45	1115.36	1312.77	1533.15	1955.41
2022_tahu	499.15	918.15	1138.43	1413.98	1573.45	2084.03
2021_tempe	446.88	813.02	1035.89	1305.45	1654.96	990.19
2022_tempe	431.73	856.11	1086.5	1416.95	1697.74	2102.11
2021_oncom	16.53	11.64	14.0	34.47	31.13	30.54
2022_oncom	16.41	10.77	15.24	37.46	31.8	33.86

Tabel 6. Interpretasi Cluster Metode Average Linkage

Pada Tabel 6 menunjukkan karakteristik masing-masing cluster. Sehingga interpretasi atau kesimpulan cluster sebagai berikut.

1. Cluster 0 dengan rata-rata variabel pada tiap objek dalam cluster terendah yaitu 4753.40. Sehingga, memiliki rata-rata pengeluaran perkapita seminggu yang terendah dibandingkan dengan kelompok lainnya. Hal ini mungkin mencerminkan tingkat konsumsi kacang yang lebih rendah
2. Cluster 1 dengan rata-rata variabel pada tiap objek yang sedikit lebih tinggi dibandingkan cluster 0 yaitu 7809.08. Hal ini menunjukkan bahwa kelompok ini pola konsumsi kacang mungkin lebih seimbang
3. Cluster 2 dengan rata-rata variabel pada tiap objek yang lebih tinggi yaitu 9651.14. Wilayah-wilayah dalam kelompok ini dapat memiliki tingkat konsumsi yang lebih tinggi atau mungkin memiliki struktur ekonomi yang mendukung pengeluaran lebih besar pada kelompok kacang-kacangan
4. Cluster 3 dengan rata-rata variabel tiap objek cukup tinggi yaitu 11852.89. Ini mungkin mencerminkan tingkat konsumsi yang tinggi atau karakteristik ekonomi yang mendukung pengeluaran lebih besar di kelompok kacang-kacangan
5. Cluster 4 menunjukkan rata-rata yang sangat tinggi yaitu 13997.48. Wilayah-wilayah dalam kelompok ini dapat diidentifikasi sebagai wilayah dengan tingkat konsumsi tertinggi atau karakteristik ekonomi yang mendukung pengeluaran besar pada kacang-kacangan
6. Cluster 5 menunjukkan rata-rata pengeluaran tinggi sekali dibandingkan rata-rata cluster yang lain yaitu 17718.31. menandakan bahwa kelompok ini mungkin memiliki tingkat konsumsi tertinggi atau struktur ekonomi yang mendukung pengeluaran yang signifikan pada kelompok kacang-kacangan.

```
jenis_kacang_tertinggi = max(zip(
    ["kacang", "kacang-tanah", "kacang-kedele", "kacang-lain", "tahu", "tempe", "oncom"],
    avg_kacang_values_2021, avg_kacang_values_2022
), key=lambda x: max(x[1], x[2]))

jenis_kacang_terendah = min(zip(
    ["kacang", "kacang-tanah", "kacang-kedele", "kacang-lain", "tahu", "tempe", "oncom"],
    avg_kacang_values_2021, avg_kacang_values_2022
), key=lambda x: max(x[1], x[2]))
```

```
# Menghitung nilai rata-rata
avg_2021 = sum(avg_kacang_values_2021) / len(avg_kacang_values_2021)
avg_2022 = sum(avg_kacang_values_2022) / len(avg_kacang_values_2022)
```

```
# Menampilkan nilai rata-rata 2021 dan 2022 dengan tiga angka di belakang koma
print(f"Nilai rata-rata 2021: {avg_2021:.3f}")
print(f"Nilai rata-rata 2022: {avg_2022:.3f}")

# Menampilkan jenis kacang dengan rata-rata tertinggi dan terendah dengan tiga angka di belakang koma
print(f"Jenis kacang dengan rata-rata tertinggi di tahun 2021 dan 2022: {jenis_kacang_tertinggi[0]}")
print(f"Jenis kacang dengan rata-rata terendah di tahun 2021 dan 2022: {jenis_kacang_terendah[0]}")
```

Gambar 3. Menampilkan Antar Tahun Rata-Rata

Pada gambar 1 didapatkan gambaran umum mengenai data pengeluaran seminggu kelompok perkabupaten/kota di Indonesia tahun 2021 sampai 2022, yaitu antar tahun rata-rata pengeluaran konsumsi kacang.

Nilai Rataan		Jenis Kacang 2021-2022	
2021	**2022**	**Tertinggi**	**Terendah**
1151.256	1384.391	Kacang	Kacang Kedele

Tabel 7. Antar Tahun Rata Rata Pengeluaran Konsumsi Kacang

Dilihat dari Tabel 7 dapat dipastikan lagi untuk tingkat jenis kacang dengan rata-rata pengeluaran penduduk, berikut ini hasilnya.

```
# Menentukan jenis kacang dengan rata-rata tertinggi
jenis_kacang_tertinggi = avg_kacang_kacangan_by_prediction.select(
    "Kabupaten/Kota",
    "prediction",
    "avg_kacang",
    "avg_kacang-tanah",
    "avg_kacang-kedele",
    "avg_kacang-lain",
    "avg_tahu",
    "avg_tempe",
    "avg_oncom",
    "avg_total"
).orderBy(F.desc("avg_total")).first()

# Menentukan jenis kacang dengan rata-rata terendah
jenis_kacang_terendah = avg_kacang_kacangan_by_prediction.select(
    "Kabupaten/Kota",
    "prediction",
    "avg_kacang",
    "avg_kacang-tanah",
    "avg_kacang-kedele",
    "avg_kacang-lain",
    "avg_tahu",
    "avg_tempe",
    "avg_oncom",
    "avg_total"
).orderBy(F.asc("avg_total")).first()
```

```
# Daftar jenis kacang
jenis_kacang_list = ["avg_kacang", "avg_kacang-tanah", "avg_kacang-kedele",
                     "avg_kacang-lain", "avg_tahu", "avg_tempe", "avg_oncom"]
```

```
# Bubble sort untuk menentukan jenis kacang dengan rata-rata tertinggi
for i in range(len(jenis_kacang_list)):
    for j in range(0, len(jenis_kacang_list)-i-1):
        if jenis_kacang_tertinggi[jenis_kacang_list[j]] < jenis_kacang_tertinggi[jenis_kacang_list[j+1]]:
            jenis_kacang_list[j], jenis_kacang_list[j+1] = jenis_kacang_list[j+1], jenis_kacang_list[j]

# Menampilkan hasil rata-rata tertinggi
print("Jenis kacang dengan rata-rata tertinggi pada tahun 2021 dan 2022:")
print("Kabupaten/Kota:", jenis_kacang_tertinggi["Kabupaten/Kota"])
print("Prediction:", jenis_kacang_tertinggi["prediction"])
print("Rata-rata Total:", jenis_kacang_tertinggi["avg_total"])
print("Jenis kacang:", jenis_kacang_list[0])
print("urutan: ", jenis_kacang_list)
```

Gambar 4. Menampilkan Tingkat Rata Rata Tertinggi

```
# Bubble sort untuk menentukan jenis kacang dengan rata-rata terendah
for i in range(len(jenis_kacang_list)):
    for j in range(0, len(jenis_kacang_list)-i-1):
        if jenis_kacang_terendah[jenis_kacang_list[j]] > jenis_kacang_terendah[jenis_kacang_list[j+1]]:
            jenis_kacang_list[j], jenis_kacang_list[j+1] = jenis_kacang_list[j+1], jenis_kacang_list[j]

# Menampilkan hasil rata-rata terendah
print("\nJenis kacang dengan rata-rata terendah pada tahun 2021 dan 2022:")
print("Kabupaten/Kota:", jenis_kacang_terendah["Kabupaten/Kota"])
print("Prediction:", jenis_kacang_terendah["prediction"])
print("Rata-rata Total:", jenis_kacang_terendah["avg_total"])
print("Jenis kacang:", jenis_kacang_list[0])
print("urutan: ", jenis_kacang_list)
```

Gambar 5. Menampilkan Tingkat Rata Rata Terendah

Pada gambar 4 dan 5 didapatkan hasil yang menunjukkan urutan jenis kacang tertinggi dan terendah dari avg_total serta nama kabupaten/kota yang terkait, berikut hasilnya.

Jenis Kacang Tertinggi 2021-2022		**Jenis Kacang Terendah 2021-2022**	
Jenis	Kacang	**Jenis**	Kacang Kedele
Kabupaten/Kota	Puncak Jaya	**Kabupaten/Kota**	Pulau Taliabu
Prediction	5	**Prediction**	0
Rata-rata	17392.515	**Rata-rata**	572.355
Urutan	• Kacang-kacangan • Kacang Tanah • Tahu • Tempe • Kacang Kedele • Kacang Lainnya • Oncom	**Urutan**	• Kacang Kedele • Oncom • Kacang Tanah • Kacang Lainnya • Tempe • Tahu • Kacang-kacangan

Tabel 8. Tingkat Rata-Rata 2021-2022

```
nilai_rata_pertahun = avg_growth_by_cluster

# Menentukan jenis kacang dengan rata-rata tertinggi 2021
jenis_tertinggi_2021 = nilai_rata_pertahun.select(
    "Kabupaten/Kota",
    "prediction",
    "avg_kacang_2021",
    "avg_kacang-tanah_2021",
    "avg_kacang-kedele_2021",
    "avg_kacang-lain_2021",
    "avg_tahu_2021",
    "avg_tempe_2021",
    "avg_oncom_2021"
).orderBy(F.col("avg_kacang_2021").desc(),
    F.col("avg_kacang-tanah_2021").desc(),
    F.col("avg_kacang-kedele_2021").desc(),
    F.col("avg_kacang-lain_2021").desc(),
    F.col("avg_tahu_2021").desc(),
    F.col("avg_tempe_2021").desc(),
    F.col("avg_oncom_2021").desc()
).first()

# Menentukan jenis kacang dengan rata-rata terendah 2021
jenis_terendah_2021 = nilai_rata_pertahun.select(
    "Kabupaten/Kota",
    "prediction",
    "avg_kacang_2021",
    "avg_kacang-tanah_2021",
    "avg_kacang-kedele_2021",
    "avg_kacang-lain_2021",
    "avg_tahu_2021",
    "avg_tempe_2021",
    "avg_oncom_2021",
).orderBy(F.col("avg_kacang_2021").asc(),
    F.col("avg_kacang-tanah_2021").asc(),
    F.col("avg_kacang-kedele_2021").asc(),
    F.col("avg_kacang-lain_2021").asc(),
    F.col("avg_tahu_2021").asc(),
    F.col("avg_tempe_2021").asc(),
    F.col("avg_oncom_2021").asc()).first()
```

```
# Daftar jenis kacang

jenis_list_tertinggi_2021 = ["avg_kacang_2021", "avg_kacang-tanah_2021", "avg_kacang-kedele_2021", "avg_kacang-lain_2021",
                  "avg_tahu_2021", "avg_tempe_2021", "avg_oncom_2021"]

jenis_list_terendah_2021 = ["avg_kacang_2021", "avg_kacang-tanah_2021", "avg_kacang-kedele_2021", "avg_kacang-lain_2021",
                  "avg_tahu_2021", "avg_tempe_2021", "avg_oncom_2021"]

jenis_list_tertinggi_2022 = ["avg_kacang_2022", "avg_kacang-tanah_2022", "avg_kacang-kedele_2022", "avg_kacang-lain_2022",
                  "avg_tahu_2022", "avg_tempe_2022", "avg_oncom_2022"]

jenis_list_terendah_2022 = ["avg_kacang_2022", "avg_kacang-tanah_2022", "avg_kacang-kedele_2022", "avg_kacang-lain_2022",
                  "avg_tahu_2022", "avg_tempe_2022", "avg_oncom_2022"]
```

```
# Bubble sort untuk menentukan jenis kacang dengan rata-rata tertinggi pada tahun 2021
for i in range(len(jenis_list_tertinggi_2021)):
    for j in range(0, len(jenis_list_tertinggi_2021)-i-1):
        if jenis_tertinggi_2021[jenis_list_tertinggi_2021[j]] < jenis_tertinggi_2021[jenis_list_tertinggi_2021[j+1]]:
            jenis_list_tertinggi_2021[j], jenis_list_tertinggi_2021[j+1] = jenis_list_tertinggi_2021[j+1], jenis_list_tertinggi_

# Bubble sort untuk menentukan jenis kacang dengan rata-rata terendah pada tahun 2021
for i in range(len(jenis_list_terendah_2021)):
    for j in range(0, len(jenis_list_terendah_2021)-i-1):
        if jenis_terendah_2021[jenis_list_terendah_2021[j]] > jenis_terendah_2021[jenis_list_terendah_2021[j+1]]:
            jenis_list_terendah_2021[j], jenis_list_terendah_2021[j+1] = jenis_list_terendah_2021[j+1], jenis_list_terendah_2021

# Menampilkan hasil jenis kacang dengan rata-rata tertinggi pada tahun 2021
print("\nJenis kacang dengan rata-rata tertinggi pada tahun 2021:")
print("Kabupaten/Kota:", jenis_tertinggi_2021["Kabupaten/Kota"])
print("Prediction:", jenis_tertinggi_2021["prediction"])
print("Jenis kacang:", jenis_list_tertinggi_2021[0])
print("urutan: ", jenis_list_tertinggi_2021)

# Menampilkan hasil jenis kacang dengan rata-rata terendah pada tahun 2021
print("\nJenis kacang dengan rata-rata terendah pada tahun 2021:")
print("Kabupaten/Kota:", jenis_terendah_2021["Kabupaten/Kota"])
print("Prediction:", jenis_terendah_2021["prediction"])
print("Jenis kacang:", jenis_list_terendah_2021[0])
print("urutan: ", jenis_list_terendah_2021)
```

Gambar 6. Menampilkan Tingkat Tahun 2021

Pada gambar 6 didapatkan hasil yang menunjukkan urutan jenis kacang tertinggi dan terendah serta nama kabupaten/kota yang terkait, berikut hasilnya.

Jenis Kacang Tertinggi 2021		**Jenis Kacang Terendah 2021**	
Jenis	Kacang	**Jenis**	Kacang Kedele
Kabupaten/Kota	Nganjuk	**Kabupaten/Kota**	Lanny Jaya
Prediction	5	**Prediction**	0
Urutan	• Kacang-kacangan • Tempe • Tahu • Kacang Tanah • Kacang Lainnya • Oncom • Kacang Kedele	**Urutan**	• Kacang Kedele • Oncom • Kacang Lainnya • Tahu • Kacang-kacang • Kacang Tanah • Tempe

Table 9. Tingkat Rata-Rata Tinggi Rendah 2021

```
# Menentukan jenis kacang dengan rata-rata tertinggi 2022
jenis_tertinggi_2022 = nilai_rata_pertahun.select(
    "Kabupaten/Kota",
    "prediction",
    "avg_kacang_2022",
    "avg_kacang-tanah_2022",
    "avg_kacang-kedele_2022",
    "avg_kacang-lain_2022",
    "avg_tahu_2022",
    "avg_tempe_2022",
    "avg_oncom_2022"
).orderBy(F.col("avg_kacang_2022").desc(),
    F.col("avg_kacang-tanah_2022").desc(),
    F.col("avg_kacang-kedele_2022").desc(),
    F.col("avg_kacang-lain_2022").desc(),
    F.col("avg_tahu_2022").desc(),
    F.col("avg_tempe_2022").desc(),
    F.col("avg_oncom_2022").desc()
).first()

# Menentukan jenis kacang dengan rata-rata terendah
jenis_terendah_2022 = nilai_rata_pertahun.select(
    "Kabupaten/Kota",
    "prediction",
    "avg_kacang_2022",
    "avg_kacang-tanah_2022",
    "avg_kacang-kedele_2022",
    "avg_kacang-lain_2022",
    "avg_tahu_2022",
    "avg_tempe_2022",
    "avg_oncom_2022",
).orderBy(F.col("avg_kacang_2022").asc(),
    F.col("avg_kacang-tanah_2022").asc(),
    F.col("avg_kacang-kedele_2022").asc(),
    F.col("avg_kacang-lain_2022").asc(),
    F.col("avg_tahu_2022").asc(),
    F.col("avg_tempe_2022").asc(),
    F.col("avg_oncom_2022").asc()).first()
```

```
# Bubble sort untuk menentukan jenis kacang dengan rata-rata terendah pada tahun 2022
for i in range(len(jenis_list_terendah_2022)):
    for j in range(0, len(jenis_list_terendah_2022)-i-1):
        if jenis_terendah_2022[jenis_list_terendah_2022[j]] > jenis_terendah_2022[jenis_list_terendah_2022[j+1]]:
            jenis_list_terendah_2022[j], jenis_list_terendah_2022[j+1] = jenis_list_terendah_2022[j+1], jenis_list_terendah_2022


# Bubble sort untuk menentukan jenis kacang dengan rata-rata tertinggi pada tahun 2022
for i in range(len(jenis_list_tertinggi_2022)):
    for j in range(0, len(jenis_list_tertinggi_2022)-i-1):
        if jenis_tertinggi_2022[jenis_list_tertinggi_2022[j]] < jenis_tertinggi_2022[jenis_list_tertinggi_2022[j+1]]:
            jenis_list_tertinggi_2022[j], jenis_list_tertinggi_2022[j+1] = jenis_list_tertinggi_2022[j+1], jenis_list_tertinggi_

# Menampilkan hasil jenis kacang dengan rata-rata tertinggi pada tahun 2022
print("\nJenis kacang dengan rata-rata tertinggi pada tahun 2022:")
print("Kabupaten/Kota:", jenis_tertinggi_2022["Kabupaten/Kota"])
print("Prediction:", jenis_tertinggi_2022["prediction"])
print("Jenis kacang:", jenis_list_tertinggi_2022[0])
print("urutan: ", jenis_list_tertinggi_2022)

# Menampilkan hasil jenis kacang dengan rata-rata terendah pada tahun 2022
print("\nJenis kacang dengan rata-rata terendah pada tahun 2022:")
print("Kabupaten/Kota:", jenis_terendah_2022["Kabupaten/Kota"])
print("Prediction:", jenis_terendah_2022["prediction"])
print("Jenis kacang:", jenis_list_terendah_2022[0])
print("urutan: ", jenis_list_terendah_2022)
```

Gambar 7. Menampilkan Tingkat Tahun 2022

Pada gambar 7 didapatkan hasil yang menunjukkan urutan jenis kacang tertinggi dan terendah serta nama kabupaten/kota yang terkait, berikut hasilnya.

Jenis Kacang Tertinggi 2022		Jenis Kacang Terendah 2022	
Jenis	Kacang	**Jenis**	Kacang Kedele
Kabupaten/Kota	Puncak Jaya	**Kabupaten/Kota**	Puncak
Prediction	5	**Prediction**	0
Urutan	• Kacang-kacangan • Kacang Tanah • Tahu • Tempe • Kacang Tanah • Kacang Lainnya • Oncom	**Urutan**	• Kacang Kedele • Tahu • Oncom • Kacang Lainnya • kacang Tanah • Tempe • Kacang-Kacangan

Table 10. Tingkat Rata-Rata Tinggi Rendah 2022

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian pada analisis cluster menggunakan algoritma *hierarchical* untuk pengelompokan wilayah-wilayah berdasarkan rata-rata pengeluaran perkapita seminggu pada kacang-kacangan kabupaten/kota tahun 2021-2022 diperoleh bahwa terdapat 6 cluster yang memiliki kemiripan karakteristik berdasarkan 14 variabel yang digunakan. Serta gambaran umum tentang analisis.

1. Cluster 0 dengan rata-rata variabel pada tiap objek dalam cluster terendah yaitu 4753.40. Sehingga, memiliki rata-rata pengeluaran perkapita seminggu yang terendah dibandingkan dengan kelompok lainnya. Hal ini mungkin mencerminkan tingkat konsumsi kacang yang lebih rendah
2. Cluster 1 dengan rata-rata variabel pada tiap objek yang sedikit lebih tinggi dibandingkan cluster 0 yaitu 7809.08. Hal ini menunjukkan bahwa kelompok ini pola konsumsi kacang mungkin lebih seimbang
3. Cluster 2 dengan rata-rata variabel pada tiap objek yang lebih tinggi yaitu 9651.14. Wilayah-wilayah dalam kelompok ini dapat memiliki tingkat konsumsi yang lebih tinggi atau mungkin memiliki struktur ekonomi yang mendukung pengeluaran lebih besar pada kelompok kacang-kacangan
4. Cluster 3 dengan rata-rata variabel tiap objek cukup tinggi yaitu 11852.89. Ini mungkin mencerminkan tingkat konsumsi yang tinggi atau karakteristik ekonomi yang mendukung pengeluaran lebih besar di kelompok kacang-kacangan
5. Cluster 4 menunjukkan rata-rata yang sangat tinggi yaitu 13997.48. Wilayah-wilayah dalam kelompok ini dapat diidentifikasi sebagai wilayah dengan tingkat konsumsi tertinggi atau karakteristik ekonomi yang mendukung pengeluaran besar pada kacang-kacangan
6. Cluster 5 menunjukkan rata-rata pengeluaran tinggi sekali dibandingkan rata-rata cluster yang lain yaitu 17718.31. menandakan bahwa kelompok ini mungkin memiliki tingkat konsumsi tertinggi atau struktur ekonomi yang mendukung pengeluaran yang signifikan pada kelompok kacang-kacangan.
7. Nilai rata-rata pada tahun 2021 sebesar 1151.256, serta nilai rata-rata tahun 2022 sebesar 1384.391
8. Jenis kacang tertinggi baik pada tahun 2021 atau 2022 yaitu jenis kacang-kacangan.
9. Jenis kacang terendah baik pada tahun 2021 atau 2022 yaitu jenis kacang kedele.
10. Didapatkan jenis kacang tertinggi pada tahun 2021-2022 adalah Kacang dengan prediction cluster ke 5 yaitu di kota/kabupaten Puncak Jaya dengan rata rata 17392.515
11. Didapatkan jenis kacang terendah pada tahun 2021-2022 adalah kacang kedele dengan prediction cluster ke 0 yaitu di kota/kabupaten Pulau Taliabu dengan rata rata 572.355
12. Didapatkan urutan jenis kacang tertinggi pada tahun 2021-2022 adalah kacang-kacangan, kacang tanah, tahu, tempe, kacang kedele, kacang lainnya, oncom.
13. Didapatkan urutan jenis kacang terendah pada tahun 2021-2022 adalah kacang kedele, oncom, kacang tanah, kacang lainnya, tempe, tahu, kacang-kacangan.

14. Didapatkan jenis kacang tertinggi pada tahun 2021 adalah Kacang dengan prediction cluster ke 5 yaitu di kota/kabupaten Ngajuk dengan urutan kacang-kacangan, tempe, tahu, kacang tanah, kacag lainnya, oncom, kacang kedele.
15. Didapatkan jenis kacang terendah pada tahun 2021 adalah kacang kedele dengan prediction cluster ke 0 yaitu di kota/kabupaten Lanny Jaya dengan urutan kacang kedele, oncom, kacang lainnya, tahu, kacang-kacangan, kacang tanah, tempe.
16. Didapatkan jenis kacang tertinggi pada tahun 2022 adalah Kacang dengan prediction cluster ke 5 yaitu di kota/kabupaten Puncak Jaya dengan urutan Kacang-kacangan, Kacang Tanah, Tahu, Tempe, Kacang Tanah, Kacang Lainnya, Oncom.
17. Didapatkan jenis kacang tertinggi pada tahun 2022 adalah kacang kedele dengan prediction cluster ke 0 yaitu di kota/kabupaten Puncak dengan urutan Kacang Kedele, Tahu, Oncom, Kacang Lainnya, kacang Tanah, Tempe, Kacang-Kacangan.

5.2 Saran

Berdasarkan hasil penelitian dan pembahasan yang mengacu pada Batasan penelitian ini, maka dapat disarankan bahwa:

1. Bagi peneliti perlunya kajian secara lebih mendalam agar model yang dikembangkan dapat memenuhi kebutuhan. Selain itu, penelitian dapat melibatkan analisis perbandingan antara kelompok berpendapatan tinggi, sedang, dan rendah untuk memahami apakah ada perbedaan signifikan dalam preferensi pengeluaran antara kelompok-kelompok tersebut.
2. Hasil penelitian yang telah di lakukan dapat menjadi rujukan, kerangka kerja dan pedoman dalam melaksanakan pengembangan model klasifikasi lain nya.

DAFTAR PUSTAKA

[1] Nisa, Khairun., (2019), Analisis Cluster Dengan Menggunakan Metode Hierarki Untuk Pengelompokan Kecamatan Di Kabupaten Langkat Berdasarkan Indikator Kesehatan. Skripsi, 6-19.

[2] Haerani, Elin., Fadhilah Syafria., (2021), Penerapan Konsep Data Science dan Machine Learning untuk Kelangsungan dan Keberhasilan Studi Mahasiswa UIN SUSKA Riau. Skripsi, 9-12.

[3] Simamora, (2005), Analisis Multivariat Pemasaran Edisi Pertama, Jakarta: PT

Gramedia Pustaka Utama.

[4] Singh dkk. (2022). *Escalate Protein Plates from Legumes for Sustainable Human Nutrition*. Diakses 25 Desember 2023.