**LAPORAN PROJECT UJIAN AKHIR SEMESTER**

**STATISTIKA DESKRIPTIF**



**NAMA : MUKHAMAD IKHSANUDIN**

**NIM : 082011633086**

**DOSEN PENGAMPU : Drs. ETO WURYANTO, DEA.**

**196609281991021001**

**PROGRAM STUDI S1 SISTEM INFORMASI**

**FAKULTAS SAINS DAN TEKNOLOGI**

**UNIVERSITAS AIRLANGGA**

**2021**

**DAFTAR ISI**

**BAB I**

**PENDAHULUAN**

1. Latar Belakang

Pada era digital, pertumbuhan data menjamur dengan sangat cepat. Hal ini disebabkan oleh beberapa faktor yang salah satunya adalah meningkatnya penggunaan smartphone dan internet. Hal tersebut kemudian menyebabkan proses pertukaran informasi menjadi sangat cepat dan mudah. Bahkan tanpa disadari, setiap hari Anda menghasilkan banyak data dan menyebarkannya melalui media sosial, aplikasi smartphone dan mesin pencari.

Dengan berjalannya waktu, data akan menghasilkan nilai jual yang sangat tinggi. Hal ini dikarena terdapat pengolahan data mentah hingga bisa menghasilkan keterangan dan fakta yang mudah dipahami oleh banyak orang. Bahwasanya data dapat diolah dengan berbagai cara agar menghasilkan output yang bisa diintepretasikan. Pengolahan data yang dapat dilakukan kembali lagi dari kebutuhan dan tujuannya. Tujuan disini merupakan hal yang sangat mendasari dari bagaimana data itu dapat melangkah agar bisa berubah menjadi sebuah informasi.

Di abad 20 sendiri pengolahan data merupakan suatu hal yang sangat melekat bagi banyak orang, baik itu data dengan volume kecil hingga besar. Terbukti dari adanya software pengolahan data yang bertujuan untuk mempermudahkan seseorang dalam mengolah data mentah hinggan menjadi rangkaian informasi yang dapat digunakan oleh banyak orang. Salah satu software yang bisa digunakan untuk mengolah data dan menganalisis data yaitu R Studio dengan bahasa program R. R merupakan program komputasi statistika dan grafis. Saat ini R dikenal sebagai powerful software untuk analisis data dan Data Science. Seiring perkembangan teknologi yang terus bertransformasi membuat kemampuan R sebagai bahasa pemrograman menjadi jauh lebih luas, terlihat dari kemampuannya mencakup hampir semua lini dalam analisis data, dan mampu menyelesaikan berbagai masalah terkait pengolahan data.

**BAB II**

**LANDASAN TEORI**

1. Data

Data adalah sekumpulan keterangan atau fakta mentah berupa simbol, angka, kata-kata, atau citra, yang didapatkan melalui proses pengamatan atau pencarian ke sumber-sumber tertentu. Pendapat lain mengatakan, definisi data adalah kumpulan keterangan-keterangan atau deskripsi dasar dari suatu hal (objek atau kejadian) yang diperoleh dari hasil pengamatan (observasi) dan dapat diolah menjadi bentuk yang lebih kompleks, seperti; informasi, database, atau solusi untuk masalah tertentu. Merujuk pada pendapat para ahli bahwasanya definisi data seperti berikut ini:

1. Arikunto Suharsimi

Menurut Arikunto Suharsimi, pengertian data adalah semua fakta dan angka-angka yang dapat dijadikan bahan untuk menyusun sebuah informasi.

1. Slamet Riyadi

Menurut Slamet Riyadi, pengertian data adalah kumpulan keterangan yang diperoleh dari pengamatan dimana data bisa berupa angka-angka atau lambang-lambang.

1. Kristanto

Menurut Kristanto, pengertian data adalah suatu fakta mentah mengenai objek yang dapat mengurangi derajat ketidakpastian tentang suatu keadaan dan kejadian.

1. R dan R Studio

R sendiri sebenarnya adalah program komputasi statistika dan grafis. R memiliki kemampuan yang mencakup hampir semua lini dalam analisis data. Penamaan “R” ini sendiri diambil dari huruf depan kedua pembuatnya, yakni Ross Ihaka dan Robert Gentleman yang membuatnya di Universitas Auckland, Selandia Baru. Bahasa pemgrogaman R menyediakan berbagai macam teknik statistika dan mampu menghasilkan sebuah grafik dengan kualitas tinggi yang dapat memuat simbol matematika, sehingga kini sudah menjadi sebuah standard internasional yang digunakan oleh semua statistikawan dalam melakukan pengembangan berbagai macam perangkat lunak yang berhubungan dengan statistika. Saat ini, R berada di bawah pengembangan GNU.

Di sisi lain, RStudio sendiri merupakan integrated development environment (IDE) khusus yang diciptakan untuk bahasa pemrograman R. Sebenarnya, tersedia dua macam RStudio, yakni open source (gratis) dan berbayar. Hal inilah yang membuat RStudio memiliki banyak pengguna karena dapat digunakan secara cuma-cuma. Selain itu, RStudio juga memiliki berbagai macam kelebihan lainnya, seperti dapat digunakan di berbagai macam platform dan fitur-fitur di dalamnya yang cukup berkualitas.

1. PCA

Principal Component Analysis (PCA) merupakan sebuah teknik pengolahan data dalam dunia statistika yang digunakan untuk melakukan reduksi dimensi data atau menyederhanakan sebuah data tanpa mengurangi makna atau karakteristik asli data sehingga nantinya akan memudahkan proses interpretasi data tersebut. Hal ini dapat dilakukan dengan mengubah variabel-variabel awal yang memiliki jumlah sebanyak x dan berkolerasi satu sama yang lain menjadi satu atau lebih variabel baru yang berisikan kombinasi dari variabel-variabel awal dengan jumlah n dan tidak lagi memiliki korelasi, yang mana jumlah n lebih sedikit daripada x. Variabel-variabel baru yang tercipta dari proses tersebutlah yang akan disebut sebagai principal component. Hasil dari principal component inilah yang nantinya akan memudahkan dalam melakukan interpretasi data karena dengan hanya menggunakan n principal component tersebut, akan didapatkan hasil yang sama dalam penginterpretasian dengan menggunakan x variabel awal. Penggunaan teknik PCA ini sangat berguna apabila data yang digunakan memiliki jumlah variabel yang cukup besar dan memiliki korelasi antar variabel-variabelnya karena tujuannya sendiri adalah mengerucutkan data menjadi sederhana sehingga lebih mudah diinterpretasikan.

1. Clustering

Clustering Data adalah sebuah teknik pengolahan data multivariate yang bertujuan untuk mengelompokkan objek-objek berdasarkan dari sifat atau karakteristik yang dimiliki. Clustering Data akan mengelompokkan objek-objek yang memiliki karakteristik yang mirip dalam satu klaster/kelompok yang sama. Jika dilogikakan, cluster yang baik adalah sebuah klaster yang memiliki kehomogenan (kemiripan) yang tinggi antar anggota dalam satu klaster dan keheterogenan (perbedaan) yang tinggi antar anggota satu klaster dengan anggota klaster lainnya. Secara garis besar, sebenarnya terdapat dua metode dalam melakukan clustering data, yakni metode hirarki dan metode non-hirarki. Metode hirarki ini memiliki konsep pengelompokkan seperti pohon yang mana pada awalnya hanya akan dikelompokkan dua atau lebih objek yang mirip kemudian baru dicari kemiripan kedua dan begitu seterusnya layaknya membentuk sebuah pohon, yang mana terlihat tingkatan yang jelas antar objek, dari yang paling mirip hingga yang paling berbeda. Pada metode hiraki, digunakan sebuah dendogram sebagai instrument penjelas untuk menunjukkan proses hirarki tersebut. Sebaliknya, metode non-hirarki tidak menunjukkan adanya tingkatan atau konsep pohon sebagaimana yang ditunjukkan pada metode hirarki. Salah satu metode non-hirarki yang seringkali digunakan adalah metode K-Means. Metode K-Means dimulai dengan menentukan terlebih dahulu jumlah klaster yang ingin dibuat kemudian baru memasukkan objek-objek ke dalam klaster yang tersedia berdasarkan dari kemiripan karakteristik yang dimiliki setiap objek. K-Means sendiri merupakan teknik yang paling lumrah digunakan jika ingin membuat sebuah Clustering Data. Selain itu juga ada metode K-Medoids, yakni sebuah varian dari metode K-Means. K-Medoids menggunakan suatu objek untuk mewakili satu cluster yang disebut dengan medoids. Setelah itu barulah dicari objek-objek non-medoids yang akan dimasukkan ke dalam cluster dengan didasari oleh kemiripan objek medoids pada setiap cluster yang ada. Terakhir, ada juga metode yang dinamakan dengan Mclust. Metode ini sendiri merupakan sebuah produk yang disediakan oleh R yang selain bisa mengelompokkan data, juga bisa melihat analisis diskriminan. Keuntungan Mclust adalah fiturnya yang terbilang cukup anyar dibanding metode lainnya karena menggunakan algoritma EM dan memaksimalkan Bayesian Information Criterion (BIC).

1. CA dan MCA

Korespondensi Analisis adalah adalah sebuah tehnik multivariate secara grafik yang digunakan untuk eksplorasi data dari sebuah tabel kontingensi. Memproyeksikan baris-baris dan kolom-kolom dari matriks data sebagai titik-titik ke dalam sebuah grafik dalam sebuah jarak Euclid.

Analisis korespondensi berganda (MCA) adalah perluasan dari analisis korespondensi sederhana untuk meringkas dan memvisualisasikan tabel data yang berisi lebih dari dua variabel kategori. Hal ini juga dapat dilihat sebagai generalisasi dari analisis komponen utama ketika variabel yang akan dianalisis bersifat kategoris daripada kuantitatif.

1. Library R

Library R merupakan tempat semua paket disimpan. Segala bentuk perhitungan secara numerik dll. Terdapat di Library R selain hitungan Library R juga memberikan plotingan visualisasi serta dapat membuat visualisasi menjadi hidup dan mudah dibaca karena terdapat informasi.

**BAB III**

**PEMAPARAN DATA**

1. PCA

Pada *Principal Component Analysis*, data yang digunakan adalah “*dentitio”*  yang berasal dari *package “flexclust”*. *Dentitio* merupakan data jumlah gigi dari berbagai mamalia yang dikelompokkan berdasarkan macamnya. Gigi mamalia dibagi menjadi 4 kelompok : gigi seri, taring, premolar dan geraham.

*Data frame* pada *dentitio* berisi variabel berikut

* top.inc (Seri atas)
* bot.inc (Seri bawah)
* top.can (Taring atas)
* bot.can (Taring bawah)
* top.pre (Premolar atas)
* bot.pre (Premolar bawah)
* top.mol (Geraham atas)
* bot.mol (Geraham bawah)

1. Clustering

Pada *Clustering*, data yang digunakan adalah “*nutrient”*  dari *package “flexclust”*. Data ini berisi pengukuran nutrisi dari berbagai makanan dari daging, ikan, atau unggas. Dataset ini terdiri atas 27 *instances* dan 5 variabel.

Data frame pada *nutrient* berisi variabel berikut.

* Energy : energi makanan (dalam kalori)
* Protein : protein pada makanan (dalam gram)
* Fat : lemak pada makanan (dalam gram)
* Calcium : kalsium pada makanan (dalam mg)
* Iron : kandungan besi pada makanan (dalam mg)

1. CA dan MCA
2. CA

Pada *Correspondence Analysis*, data yang digunakan adalah “*tithonia”* dari *package “ade4”*. Dataset *tithonia* menggambarkan filogeni dari 11 bunga seperti yang dilaporkan oleh Morales (2000). Dataset ini juga memberikan ciri-ciri morfologis dan demografis yang sesuai dengan 11 spesies ini.. Di dalamnya terdapat dua daftar (*list*) yaitu *tre* dan *tab*. *Tre* adalah string karakter yang menggambarkan pohon filogenetik dalam format Newick. Sedangkan *tab* adalah *data frame* dengan 11 spesies dan 14 sifat (6 sifat morfologi dan 8 demografi).

*Data frame* pada *tithonia* berisi variabel berikut.

* morho1: adalah vektor numerik yang menggambarkan ukuran benih (mm)
* morho2: adalah vektor numerik yang menggambarkan ukuran bunga (mm)
* morho3: adalah vektor numerik yang menggambarkan ukuran daun betina (cm)
* morho4: adalah vektor numerik yang menggambarkan ukuran kepala (mm)
* morho5: adalah vektor bilangan bulat yang menggambarkan jumlah bunga per kepala
* morho6: adalah vektor bilangan bulat yang menggambarkan jumlah biji per kepala
* demo7: adalah vektor numerik yang menggambarkan tinggi bibit (cm)
* demo8: adalah vektor numerik yang menggambarkan laju pertumbuhan (cm/hari)
* demo9: adalah vektor numerik yang menggambarkan waktu perkecambahan
* demo10: adalah vektor numerik yang menggambarkan pembentukan (persen)
* demo11: adalah vektor numerik yang menggambarkan viabilitas (persen)
* demo12: adalah vektor numerik yang menggambarkan perkecambahan (persen)
* demo13: adalah vektor bilangan bulat yang menggambarkan alokasi sumber daya
* demo14: adalah vektor numerik yang menggambarkan tinggi dewasa (m)

1. MCA

Pada *Multiple Correspondence Analysis*, data yang digunakan adalah *“quine”* dari *package “MASS”*. Dataset *quine* memiliki 146 baris dan 5 kolom yang didapatkan dari anak-anak di Walgett, New South Wales, dan Australia yang diklasifikasikan berdasarkan suku, usia, jenis kelamin dan status pelajar serta jumlah hari absen dari sekolah pada tahun ajaran tertentu.

*Data frame* pada *quine* berisi variabel berikut.

* Eth

Latar belakang suku bangsa : Aboriginal atau Not, ("A" or "N").

* Sex

Jenis kelamin ("F" or "M").

* Age

Kelompok umur: Utama ("F0"), bentuk lainnya "F1," "F2" or "F3".

* Lrn

Status pelajar : Average atau Slow Learner, ("AL" or "SL").

* Days

Jumlah ketidakhadiran (absen) dalam setahun

1. Library

library(factoextra)

library(rgl)

library(mclust)

library(flexclust)

library(cluster)

library(Gmedian)

library(FactoMineR)

library(corrplot)

library(MASS)

library(ade4)

* cluster digunakan untuk perhitungan clustering,
* mclust digunakan untuk menghitung klister optimal menggunakan metode BIC
* Gmedian digunakan untuk mencari nilai K-median
* FactoMineR digunakan untuk analisis data eksplorasi multivariat dan penambangan data
* factoextra, corrplot dan rgl digunakan dalam visualisasi olahan data
* flexclust, MASS, dan ade4 berisi dataset yang akan diolah

**BAB IV**

**ISI**

1. PCA
2. Clustering
3. CA dan MCA

**BAB V**

**PENUTUP**

1. Kesimpulan
2. Saran