Nama : Elena Ghini Rachman

NIM : 2502055204

Group8
GSLC Assignment Week 9 Data Mining and Visualization - DTSC6005001

Laporan EDA (Laporan Eksplorasi Data)

Shipping E-Commerce

(R language)

Data di ambil dari Kaggle dengan link sebagai berikut.

https://www.kaggle.com/code/hamzamanssor/shipping-e-commerce-ml-models

Link Github: https://github.com/elenaghini/Shipping_e-commerse

Harap lakukan tugas berikut untuk tugas GSLC Anda:

1. Pilih 1 variabel dependen dari kumpulan data pilihan Anda untuk AOL, dan satu set variabel independen. Hitung korelasi antara variabel dependen dan variabel independen, analisis apa yang tersirat dari nilai-nilai tersebut. Pastikan menggunakan metode korelasi yang benar sesuai dengan jenis variabel yang dibandingkan!

Jawaban

Data excel csv bernama shipping_ecommerce memiliki 10 variabel dan salah satunya memiliki variabel dependen yaitu customer_rating. Selanjutnya 9 variabel indenpenden yaitu variabel Customer_care_calls, Prior_purchases, Discount_offered, Weight_in_gms, Class, Warehouse_block, Mode_of_Shipment, Product_importance, dan Gender.

Berikut implementasi menggunakan bahasa R:

```
# Load the required libraries
install.packages("Hmisc")
library(readr)
library(Hmisc)

# Read the dataset
data <- read_csv("shipping_ecommerce.csv")

#mencari kolersi
cor(data$Customer_rating, data$Customer_care_calls, method = "pearson")
cor(data$Customer_rating, data$Discount_offered, method = "pearson")
cor(data$Customer_rating, data$Prior_purchases, method = "pearson")
cor(data$Customer_rating, data$Weight_in_gms, method = "pearson")
cor(data$Customer_rating, data$Class, method = "pearson")
```

Penjelaskan Code:

library(readr) yang memuat paket "readr" untuk membaca file CSV.

library(Hmisc) untuk menyediakan fungsi tambahan untuk analisis statistik.

Membaca file CSV "shipping_ecommerce.csv" dan menyimpannya dalam variabel "data" dengan syntax: data <- read_csv("shipping_ecommerce.csv")

cor(data$Customer_rating, data$Customer_care_calls, method = "pearson") syntax untuk menghitung korelasi Pearson antara variabel "Customer_rating" dan "Customer_care_calls".

Begitu pula dengan syntax-syntax yang lain menghitung korelasi Pearson antara variabel "Customer_rating" dan "Discount_offered". Variabel "Customer_rating" dengan "Prior_purchases", variabel "Customer_rating" dengan "Weight_in_gms" dan variabel "Customer_rating" dengan "Class".

Output:

```
> #mencari kolersi
> cor(data$Customer_rating, data$Customer_care_calls, method = "pearson")
[1] 0.01226955
>
> cor(data$Customer_rating, data$Discount_offered, method = "pearson")
[1] -0.003103001
>
> cor(data$Customer_rating, data$Prior_purchases, method = "pearson")
[1] 0.0131613
>
> cor(data$Customer_rating, data$Weight_in_gms, method = "pearson")
[1] -0.001959518
>
> cor(data$Customer_rating, data$Class, method = "pearson")
[1] 0.01319878
```

Metode korelasi Pearson untuk mengukur hubungan linier antara variabel-variabel yang berbeda. Berikut adalah interpretasi nilai-nilai korelasi yang Anda peroleh:

1. Korelasi antara Customer_rating dan Customer_care_calls
   Nilai korelasi sebesar 0.01226955 menunjukkan bahwa ada hubungan yang sangat lemah dan hampir tidak signifikan antara rating pelanggan dan jumlah panggilan layanan pelanggan. Korelasi positif menunjukkan bahwa adanya peningkatan rating pelanggan cenderung berkorelasi dengan peningkatan jumlah panggilan layanan pelanggan, tetapi hubungannya sangat rendah.

2. Korelasi antara Customer_rating dan Discount_offered
   Nilai korelasi sebesar -0.003103001 menunjukkan bahwa tidak ada hubungan yang signifikan antara rating pelanggan dan diskon yang ditawarkan. Nilai korelasi yang mendekati nol menunjukkan bahwa tidak ada hubungan linier yang jelas antara kedua variabel ini.

3. Korelasi antara Customer_rating dan Prior_purchases
   Nilai korelasi sebesar 0.0131613 menunjukkan bahwa ada hubungan yang sangat lemah dan hampir tidak signifikan antara rating pelanggan dan jumlah pembelian sebelumnya. Korelasi positif menunjukkan bahwa adanya peningkatan rating pelanggan cenderung berkorelasi dengan peningkatan jumlah pembelian sebelumnya, tetapi hubungannya sangat rendah.

4. Korelasi antara Customer_rating dan Weight_in_gms
   Nilai korelasi sebesar -0.001959518 menunjukkan bahwa tidak ada hubungan yang signifikan antara rating pelanggan dan berat barang dalam gram. Nilai korelasi yang mendekati nol menunjukkan bahwa tidak ada hubungan linier yang jelas antara kedua variabel ini.

5. Korelasi antara Customer_rating dan Class
   Nilai korelasi sebesar 0.01319878 menunjukkan bahwa ada hubungan yang sangat lemah dan hampir tidak signifikan antara rating pelanggan dan kelas barang. Korelasi positif menunjukkan

bahwa adanya peningkatan rating pelanggan cenderung berkorelasi dengan peningkatan kelas barang, tetapi hubungannya sangat rendah.

Berdasarkan hasil analisis korelasi yang telah dilakukan, kesimpulan yang dapat diambil adalah sebagai berikut:

Tidak ada hubungan yang signifikan antara rating pelanggan dengan jumlah panggilan layanan pelanggan, diskon yang ditawarkan, berat barang dalam gram, dan kelas barang. Korelasi antara rating pelanggan dan variabel-variabel tersebut sangat rendah atau mendekati nol, menunjukkan bahwa hubungan linier antara kedua variabel tersebut hampir tidak ada.

Terdapat hubungan yang sangat lemah dan hampir tidak signifikan antara rating pelanggan dengan jumlah pembelian sebelumnya. Korelasi positif yang sangat rendah menunjukkan bahwa ada kecenderungan peningkatan rating pelanggan yang berkorelasi dengan peningkatan jumlah pembelian sebelumnya, tetapi hubungannya sangat lemah.

2. Mengapa kita perlu melakukan uji statistik? Jika memungkinkan pada kumpulan data Anda, coba lakukan pengujian ini dan analisis apa artinya. Berikut adalah daftar atau artikel yang dapat membantu Anda mempelajari lebih lanjut tentang mereka.
   https://medium.com/@anushka.da3/types-of-statistical-tests-b8ceb90e13b3
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6639881/

Jawaban

Kita perlu melakukan uji statistik untuk mendapatkan pemahaman yang lebih baik tentang hubungan antara variabel independen (Customer_care_calls, Prior_purchases, Discount_offered, Weight_in_gms, Class, Warehouse_block, Mode_of_Shipment, Product_importance, dan Gender) dan variabel dependen (customer_rating) dalam dataset shipping_ecommerce. Uji statistik membantu kita dalam menguji hipotesis dan mengevaluasi signifikansi statistik dari hubungan tersebut.

Untuk melakukan pengujian, kita dapat menggunakan berbagai metode statistik, tergantung pada jenis data dan tujuan analisis yang ingin dicapai. Berikut ini beberapa contoh pengujian yang dapat dilakukan:

1. Analisis Korelasi
   Kita dapat menggunakan uji korelasi (misalnya, korelasi Pearson) untuk melihat apakah ada hubungan linier antara variabel independen dan variabel dependen. Jika terdapat korelasi yang kuat dan signifikan, maka variabel independen memiliki pengaruh terhadap variabel dependen.

2. Analisis Regresi
   Kita dapat melakukan analisis regresi (misalnya, regresi linier) untuk mengukur sejauh mana variabel independen mempengaruhi variabel dependen. Hal ini dapat membantu kita dalam membangun model prediksi untuk customer_rating berdasarkan variabel independen yang ada.

```
> # Linear regression model
> lm_model <- lm(Customer_rating ~ Prior_purchases + Discount_offered, data = data)
> summary(lm_model)

Call:
lm(formula = Customer_rating ~ Prior_purchases + Discount_offered,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.07022 -0.99768  0.01473  1.02322  2.03758

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.9497897  0.0369992  79.726   <2e-16 ***
Prior_purchases   0.0120610  0.0088821   1.358    0.175
Discount_offered -0.0001769  0.0008347  -0.212    0.832
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 10995 degrees of freedom
Multiple R-squared:  0.0001773, Adjusted R-squared:  -4.566e-06
F-statistic: 0.9749 on 2 and 10995 DF,  p-value: 0.3773

>
```

Dalam model ini, nilai p-value untuk kedua koefisien Prior_purchases (0.175) dan Discount_offered (0.832) lebih besar dari tingkat signifikansi yang umum digunakan (misalnya 0.05), sehingga tidak ada bukti yang cukup untuk menolak hipotesis nol, yang berarti bahwa kedua variabel independen tidak memiliki pengaruh yang signifikan terhadap Customer_rating. Model regresi linear ini memiliki kemampuan yang sangat rendah dalam menjelaskan variasi dalam Customer_rating dan tidak cocok untuk digunakan dalam memprediksi atau menjelaskan Customer_rating berdasarkan variabel independen yang digunakan.

3. Analisis ANOVA
   ANOVA adalah metode statistik yang digunakan untuk membandingkan mean (rata-rata) antara tiga atau lebih kelompok independen. Dalam dataset ini, jika kita ingin membandingkan mean customer_rating antara beberapa kelompok berdasarkan variabel independen seperti Warehouse_block, kita dapat menggunakan ANOVA. Hal ini akan membantu kita mengetahui apakah ada perbedaan signifikan dalam customer_rating di antara kelompok-kelompok tersebut.

```
> #anova
> anova_model <- aov(Customer_rating ~ Warehouse_block, data = data)
> summary(anova_model)
                Df Sum Sq Mean Sq F value Pr(>F)
Warehouse_block  4      4  0.9526   0.477  0.753
Residuals       10993  21972  1.9987
```

Berikut adalah hasil analisis ANOVA yang telat dijalankan:

1. Warehouse_block
   Derajat kebebasan (df) antara kelompok Warehouse_block adalah 4.
   Jumlah kuadrat antara kelompok (Sum Sq) adalah 4.
   Mean square (Mean Sq) adalah 0.9526.
   Nilai F yang diperoleh adalah 0.477.
   Nilai p-value (Pr(>F)) yang diperoleh adalah 0.753.
2. Residuals
   Derajat kebebasan (df) untuk residu adalah 10993.
   Jumlah kuadrat residu (Sum Sq) adalah 21972.
   Mean square residu (Mean Sq) adalah 1.9987.

Kesimpulannya hasil analisis ANOVA menunjukkan bahwa tidak ada perbedaan yang signifikan dalam rata-rata Customer_rating antara kelompok Warehouse_block. Nilai p-value yang tinggi (0.753) menunjukkan tidak adanya bukti yang cukup untuk menolak hipotesis nol, yang berarti tidak ada perbedaan yang signifikan dalam rata-rata Customer_rating antara kelompok Warehouse_block. Analisis ANOVA menunjukkan bahwa variabel Warehouse_block tidak memiliki pengaruh yang signifikan terhadap Customer_rating. Tidak ada perbedaan yang signifikan dalam rata-rata Customer_rating antara kelompok Warehouse_block.

4. Chi-Square
   Uji chi-square digunakan untuk menguji hubungan antara dua variabel kategorikal. Dalam data set ini, kita dapat menggunakan uji chi-square untuk menguji apakah ada hubungan antara variabel independen (misalnya, Warehouse_block, Mode_of_Shipment, Product_importance) dan variabel dependen (customer_rating). Misalnya, kita dapat menguji apakah ada hubungan antara Product_importance dan Mode_of_Shipment dan yang lainnya, contoh implementasinya sebagai berikut.

```
> #CHISQUARE
> chi_square <- chisq.test(table(data$Warehouse_block, data$Mode_of_Shipment))
> chi_square

        Pearson's Chi-squared test

data:  table(data$Warehouse_block, data$Mode_of_Shipment)
X-squared = 0.027478, df = 8, p-value = 1

> chi_square1 <- chisq.test(table(data$Mode_of_Shipment, data$Product_importance))
> chi_square1

        Pearson's Chi-squared test

data:  table(data$Mode_of_Shipment, data$Product_importance)
X-squared = 2.2139, df = 4, p-value = 0.6965

> chi_square2 <- chisq.test(table(data$Gender, data$Product_importance))
> chi_square2

        Pearson's Chi-squared test

data:  table(data$Gender, data$Product_importance)
X-squared = 1.1183, df = 2, p-value = 0.5717

> chi_square3 <- chisq.test(table(data$Warehouse_block, data$Product_importance))
> chi_square3

        Pearson's Chi-squared test

data:  table(data$Warehouse_block, data$Product_importance)
X-squared = 8, df = 8, p-value = 0.4335
```

Dalam analisis Chi-square yang kita lakukan, kita menguji hubungan antara beberapa variabel kategorikal dalam dataset Anda. Berikut adalah hasil analisis Chi-square yang Anda jalankan:

1. **Chi-square test antara Warehouse_block dan Mode_of_Shipment**
   Statistik Chi-square (X-squared) yang diperoleh adalah 0.027478.
   Derajat kebebasan (df) adalah 8.
   Nilai p-value yang diperoleh adalah 1.
   Kesimpulannya tidak ada hubungan yang signifikan antara variabel Warehouse_block dan Mode_of_Shipment. Nilai p-value yang tinggi (1) menunjukkan tidak adanya bukti yang cukup untuk menolak hipotesis nol, yang berarti tidak ada hubungan yang signifikan antara kedua variabel tersebut.

2. **Chi-square test antara Mode_of_Shipment dan Product_importance**
   Statistik Chi-square (X-squared) yang diperoleh adalah 2.2139.
   Derajat kebebasan (df) adalah 4.
   Nilai p-value yang diperoleh adalah 0.6965.
   Kesimpulannya tidak ada hubungan yang signifikan antara variabel Mode_of_Shipment dan Product_importance. Nilai p-value yang tinggi (0.6965) menunjukkan tidak adanya bukti yang cukup untuk menolak hipotesis nol, yang berarti tidak ada hubungan yang signifikan antara kedua variabel tersebut.

3. **Chi-square test antara Gender dan Product_importance**
   Statistik Chi-square (X-squared) yang diperoleh adalah 1.1183.

Derajat kebebasan (df) adalah 2.
Nilai p-value yang diperoleh adalah 0.5717.
Kesimpulannya tidak ada hubungan yang signifikan antara variabel Gender dan Product_importance. Nilai p-value yang tinggi (0.5717) menunjukkan tidak adanya bukti yang cukup untuk menolak hipotesis nol, yang berarti tidak ada hubungan yang signifikan antara kedua variabel tersebut.

4. **Chi-square test antara Warehouse_block dan Product_importance**
   Statistik Chi-square (X-squared) yang diperoleh adalah 8.
   Derajat kebebasan (df) adalah 8.
   Nilai p-value yang diperoleh adalah 0.4335.
   Kesimpulannya tidak ada hubungan yang signifikan antara variabel Warehouse_block dan Product_importance. Nilai p-value yang tinggi (0.4335) menunjukkan tidak adanya bukti yang cukup untuk menolak hipotesis nol, yang berarti tidak ada hubungan yang signifikan antara kedua variabel tersebut.

Dalam kesimpulannya, analisis Chi-square menunjukkan bahwa tidak ada hubungan yang signifikan antara variabel-variabel yang diuji dalam dataset Anda. Nilai p-value yang tinggi menunjukkan tidak adanya bukti yang cukup untuk menolak hipotesis nol, yang berarti tidak ada hubungan yang signifikan antara variabel-variabel tersebut.

Setelah melakukan pengujian, kita dapat melihat hasilnya untuk mengambil kesimpulan. Jika terdapat hubungan yang signifikan antara variabel independen dan variabel dependen, maka variabel independen tersebut memiliki pengaruh terhadap customer_rating dalam konteks dataset shipping_ecommerce. Namun, jika tidak ada hubungan yang signifikan, maka variabel independen mungkin tidak memiliki pengaruh yang kuat terhadap customer_rating atau terdapat faktor-faktor lain yang perlu dipertimbangkan.

5. T-test

Metode t-test digunakan untuk membandingkan mean (rata-rata) dari dua kelompok yang independen. Dalam dataset ini, kita dapat menggunakan t-test untuk membandingkan mean customer_rating antara dua kelompok yang berbeda. Misalnya, kita dapat membandingkan mean customer_rating antara kelompok berbeda berdasarkan variabel Gender. Hal ini akan membantu kita mengetahui apakah ada perbedaan signifikan dalam customer_rating antara kelompok gender tertentu.

```
> male_ratings <- data$Customer_rating[data$Gender == "M"]
> female_ratings <- data$Customer_rating[data$Gender == "F"]
> t_test <- t.test(male_ratings, female_ratings)
> t_test

        Welch Two Sample t-test

data:  male_ratings and female_ratings
t = 0.29776, df = 10995, p-value = 0.7659
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04481952  0.06087470
sample estimates:
mean of x mean of y
 2.994499  2.986472
```

Dua sampel untuk membandingkan rata-rata dari dua kelompok data yang berbeda, yaitu male_ratings dan female_ratings.

Hasil pengujian menunjukkan nilai t-statistik sebesar 0.29776 dengan derajat kebebasan (df) sebesar 10995. Nilai p-value yang dihasilkan adalah 0.7659.

Pada tingkat signifikansi 0.05, jika p-value lebih kecil dari 0.05, maka kita dapat menolak hipotesis nol (null hypothesis) bahwa tidak ada perbedaan yang signifikan antara rata-rata kedua kelompok. Namun, dalam kasus ini, p-value (0.7659) lebih besar dari 0.05, sehingga tidak ada cukup bukti statistik untuk menolak

hipotesis nol. Oleh karena itu, dapat dikatakan bahwa tidak ada perbedaan yang signifikan antara rata-rata male_ratings dan female_ratings.

Selain itu, interval kepercayaan 95% untuk perbedaan rata-rata antara kedua kelompok adalah dari -0.04481952 hingga 0.06087470. Hal ini menunjukkan rentang perkiraan di mana perbedaan sebenarnya antara rata-rata kedua kelompok mungkin berada.

Estimasi rata-rata untuk male_ratings adalah 2.994499, sementara estimasi rata-rata untuk female_ratings adalah 2.986472.

Implementasi lain:

```
> class1<- data$Customer_rating[data$Class==0]
> class2<- data$Customer_rating[data$Class==1]
> t_test1<-t.test(class1, class2)
> t_tes1
Error: object 't_tes1' not found
> t_test1

        Welch Two Sample t-test

data:  class1 and class2
t = -1.3838, df = 9505.5, p-value = 0.1665
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09191156  0.01584445
sample estimates:
mean of x mean of y
 2.967756  3.005790
```

Tujuannya adalah untuk membandingkan rata-rata dari dua kelompok data yang berbeda, yaitu class1 dan class2.

Hasil pengujian menunjukkan nilai t-statistik sebesar -1.3838 dengan derajat kebebasan (df) sebesar 9505.5. Nilai p-value yang dihasilkan adalah 0.1665.

Pada tingkat signifikansi 0.05, jika p-value lebih kecil dari 0.05, maka kita dapat menolak hipotesis nol (null hypothesis) bahwa tidak ada perbedaan yang signifikan antara rata-rata kedua kelompok. Namun, dalam kasus ini, p-value (0.1665) lebih besar dari 0.05, sehingga tidak ada cukup bukti statistik untuk menolak hipotesis nol. Oleh karena itu, dapat dikatakan bahwa tidak ada perbedaan yang signifikan antara rata-rata class1 dan class2.

Selain itu, interval kepercayaan 95% untuk perbedaan rata-rata antara kedua kelompok adalah dari -0.09191156 hingga 0.01584445. Hal ini menunjukkan rentang perkiraan di mana perbedaan sebenarnya antara rata-rata kedua kelompok mungkin berada.

Estimasi rata-rata untuk class1 adalah 2.967756, sementara estimasi rata-rata untuk class2 adalah 3.005790.

Uji statistic lain dengan summary(data):

```
> summary(data)
 Customer_care_calls Customer_rating Prior_purchases  Discount_offered Weight_in_gms
 Min.   :2.000       Min.   :1.00    Min.   : 2.000   Min.   : 1.00    Min.   :1001
 1st Qu.:3.000       1st Qu.:2.00    1st Qu.: 3.000   1st Qu.: 4.00    1st Qu.:1839
 Median :4.000       Median :3.00    Median : 3.000   Median : 7.00    Median :4149
 Mean   :4.055       Mean   :2.99    Mean   : 3.568   Mean   :13.37    Mean   :3634
 3rd Qu.:5.000       3rd Qu.:4.00    3rd Qu.: 4.000   3rd Qu.:10.00    3rd Qu.:5050
 Max.   :7.000       Max.   :5.00    Max.   :10.000   Max.   :65.00    Max.   :7846
 Warehouse_block    Mode_of_Shipment   Product_importance    Gender              Class
 Length:10998       Length:10998       Length:10998       Length:10998       Min.   :0.0000
 Class :character   Class :character   Class :character   Class :character   1st Qu.:0.0000
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :1.0000
                                                                             Mean   :0.5967
                                                                             3rd Qu.:1.0000
                                                                             Max.   :1.0000
```

3. Latih model pembelajaran mesin prediktif di R untuk memprediksi variabel dependen pilihan Anda! Ingat, harap minta semua anggota grup Anda mencoba model prediksi yang berbeda (mis. Satu menggunakan Random Forest, yang lain menggunakan Support Vector Machine, sementara yang lain menggunakan Multi-layer Perceptron). Evaluasi model dan jelaskan hasilnya!

Harap kirimkan kode R Anda (penurunan harga R atau R, keduanya dapat diterima. Pastikan untuk menulis beberapa dokumentasi untuk menjelaskan apa yang dilakukan setiap baris kode) dan laporan EDA dalam format .pdf (Anda dapat menggunakan kata, powerpoint, canva, dll.)! Anda disarankan untuk mengunggah kode ke Github dan memasukkan tautannya ke dalam laporan, sehingga Anda dapat menerbitkannya nanti sebagai bagian dari portofolio Anda selama Anda menjadi siswa. Jangan ragu untuk menghubungi saya jika Anda memiliki pertanyaan. Terima kasih dan GBU semuanya.

Catatan tambahan: JANGAN COPY DAN PASTE LAPORAN DARI TEMAN ANDA! Tulis versi Anda sendiri! Plagiarisme tidak akan ditoleransi!

Jawaban

Menggunakan Random Forest

```
> library(caret)
> library(randomForest)
> data <- read.csv("shipping_ecommerce.csv")
> # Preprocess the data
> data$Customer_rating <- as.factor(data$Customer_rating)
> data$Customer_care_calls <- as.numeric(data$Customer_care_calls)
> data$Prior_purchases <- as.numeric(data$Prior_purchases)
> data$Discount_offered <- as.numeric(data$Discount_offered)
> data$Weight_in_gms <- as.numeric(data$Weight_in_gms)
> data$Warehouse_block <- as.factor(data$Warehouse_block)
> data$Mode_of_Shipment <- as.factor(data$Mode_of_Shipment)
> data$Product_importance <- as.factor(data$Product_importance)
> data$Gender <- as.factor(data$Gender)
> data$Class <- as.factor(data$Class)
> # Split the data into training and testing sets
> set.seed(123)
> trainIndex <- createDataPartition(data$Customer_rating, p = 0.1, list = FALSE)
> trainData <- data[trainIndex, ]
> testData <- data[-trainIndex, ]
> # Train the random forest model
> model <- randomForest(Customer_rating ~ Customer_care_calls + Prior_purchases
+ Discount_offered + Weight_in_gms + Warehouse_block + Mode_of_Shipment + Produc
t_importance + Gender + Class, data = trainData)
> # Make predictions on the test data
> predictions <- predict(model, newdata = testData)
> # Evaluate the model
> confusionMatrix(predictions, testData$Customer_rating)
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5
         1 429 345 383 375 397
         2 443 407 431 378 396
         3 411 415 440 425 414
         4 367 407 396 385 365
         5 361 374 365 406 381

Overall Statistics

               Accuracy : 0.2063
                 95% CI : (0.1984, 0.2145)
    No Information Rate : 0.2036
    P-Value [Acc > NIR] : 0.253694

                  Kappa : 0.0079
```

```
 Mcnemar's Test P-Value : 0.009825

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity           0.21333  0.20893  0.21836   0.1955   0.1951
Specificity           0.80977  0.79265  0.78873   0.8064   0.8104
Pos Pred Value        0.22240  0.19805  0.20903   0.2005   0.2019
Neg Pred Value        0.80143  0.80347  0.79784   0.8014   0.8037
Prevalence            0.20321  0.19685  0.20362   0.1990   0.1974
Detection Rate        0.04335  0.04113  0.04446   0.0389   0.0385
Detection Prevalence  0.19493  0.20766  0.21271   0.1940   0.1907
Balanced Accuracy     0.51155  0.50079  0.50355   0.5009   0.5027
```

Berikut adalah penjelasan dari setiap baris kode yang digunakan dalam latihan ini:

library(caret) untuk memuat paket "caret" yang digunakan untuk pemrosesan data dan evaluasi model.

library(randomForest) untuk memuat paket "randomForest" yang digunakan untuk membangun model Random Forest.

data <- read.csv("shipping_ecommerce.csv") untuk membaca file CSV dengan nama "shipping_ecommerce.csv" dan menyimpannya ke dalam objek "data".

data$Customer_rating <- as.factor(data$Customer_rating) untuk mengubah variabel "Customer_rating" menjadi faktor.

data$Customer_care_calls <- as.numeric(data$Customer_care_calls) untuk mengubah variabel "Customer_care_calls" menjadi numerik.

data$Prior_purchases <- as.numeric(data$Prior_purchases) untuk mengubah variabel "Prior_purchases" menjadi numerik.

data$Discount_offered <- as.numeric(data$Discount_offered) untuk mengubah variabel "Discount_offered" menjadi numerik.

data$Weight_in_gms <- as.numeric(data$Weight_in_gms) untuk mengubah variabel "Weight_in_gms" menjadi numerik.

data$Warehouse_block <- as.factor(data$Warehouse_block) untuk mengubah variabel "Warehouse_block" menjadi faktor.

data$Mode_of_Shipment <- as.factor(data$Mode_of_Shipment) untuk mengubah variabel "Mode_of_Shipment" menjadi faktor.

data$Product_importance <- as.factor(data$Product_importance) untuk mengubah variabel "Product_importance" menjadi faktor.

data$Gender <- as.factor(data$Gender) untuk mengubah variabel "Gender" menjadi faktor.

data$Class <- as.factor(data$Class) untuk mengubah variabel "Class" menjadi faktor.

set.seed(123) untuk Mengatur biji acak untuk memastikan reproduktibilitas.

trainIndex <- createDataPartition(data$Customer_rating, p = 0.1, list = FALSE) untuk membuat partisi data dengan membagi variabel "Customer_rating" menjadi set pelatihan (90%) dan set pengujian (10%).

trainData <- data[trainIndex,] untuk mengambil data pelatihan berdasarkan indeks yang dihasilkan sebelumnya.

testData <- data[-trainIndex,]untuk mengambil data pengujian yang tidak termasuk dalam indeks data pelatihan.

model <- randomForest(Customer_rating ~ Customer_care_calls + Prior_purchases + Discount_offered + Weight_in_gms + Warehouse_block + Mode_of_Shipment + Product_importance + Gender + Class, data = trainData) untuk melatih model Random Forest dengan variabel dependen "Customer_rating" dan variabel prediktor yang terdaftar.

predictions <- predict(model, newdata = testData) untuk membuat prediksi menggunakan model yang telah dilatih pada data pengujian.

confusionMatrix(predictions, testData$Customer_rating) untuk mengevaluasi model dengan membuat matriks kebingungan (confusion matrix) antara prediksi dan nilai sebenarnya dari variabel "Customer_rating".

Hasil evaluasi model menunjukkan statistik kebingungan (confusion matrix) yang meliputi akurasi, sensitivitas, spesifisitas, nilai prediksi positif (Pos Pred Value), nilai prediksi negatif (Neg Pred Value), prevalensi, dan tingkat deteksi. Berikut adalah penjelasan lebih rinci mengenai hasil evaluasi model:

Confusion Matrix and Statistics adalah matriks kebingungan yang menunjukkan jumlah prediksi yang benar dan salah untuk setiap kelas prediksi dan nilai sebenarnya.

Overall Statistics:

1. Accuracy
   Akurasi model, yaitu persentase prediksi yang benar secara keseluruhan. Dalam kasus ini, akurasi adalah 0.2063 atau sekitar 20.63%. Sehingga persentase tersebut akurasinya sangatlah rendah.
2. 95% CI: Interval kepercayaan 95% untuk akurasi model.
3. No Information Rate (NIR)
4. Tingkat akurasi jika hanya menggunakan nilai sebenarnya yang paling umum dalam data pelatihan. NIR dalam kasus ini adalah 0.2036 atau sekitar 20.36%.
5. P-Value [Acc > NIR]
   Nilai p-value untuk menguji apakah akurasi model lebih baik daripada NIR. Dalam kasus ini, p-value adalah 0.253694, yang menunjukkan bahwa tidak ada cukup bukti untuk menyimpulkan bahwa model memiliki akurasi yang signifikan dibandingkan dengan NIR.
6. Kappa
   Koefisien kappa yang mengukur tingkat kesepakatan antara prediksi model dan nilai sebenarnya. Dalam kasus ini, kappa adalah 0.0079, yang menunjukkan tingkat kesepakatan yang sangat rendah.

Statistics by Class yaitu Statistik evaluasi untuk setiap kelas prediksi:

- Sensitivity
  Juga dikenal sebagai True Positive Rate atau Recall, ini mengukur proporsi positif yang benar diprediksi oleh model untuk setiap kelas. Nilai sensitivitas yang lebih tinggi menunjukkan bahwa model lebih baik dalam mendeteksi kasus positif. Namun, dalam kasus ini, sensitivitas untuk semua kelas sangat rendah, berkisar antara 0.1951 hingga 0.21836.
- Specificity
  Juga dikenal sebagai True Negative Rate, ini mengukur proporsi negatif yang benar diprediksi oleh model untuk setiap kelas. Nilai spesifisitas yang lebih tinggi menunjukkan bahwa model lebih baik dalam membedakan kasus negatif. Dalam kasus ini, spesifisitas untuk semua kelas berada dalam kisaran 0.78873 hingga 0.8104.
- Pos Pred Value
  Juga dikenal sebagai Precision, ini mengukur proporsi prediksi positif yang benar dibandingkan dengan total prediksi positif untuk setiap kelas. Nilai Pos Pred Value yang lebih tinggi menunjukkan bahwa model lebih baik dalam memberikan prediksi positif yang akurat. Namun, dalam kasus ini, nilai Pos Pred Value untuk semua kelas relatif rendah, berkisar antara 0.19805 hingga 0.22240.
- Neg Pred Value

Ini mengukur proporsi prediksi negatif yang benar dibandingkan dengan total prediksi negatif untuk setiap kelas. Nilai Neg Pred Value yang lebih tinggi menunjukkan bahwa model lebih baik dalam memberikan prediksi negatif yang akurat. Dalam kasus ini, nilai Neg Pred Value untuk semua kelas berada dalam kisaran 0.79784 hingga 0.8037.

- Prevalence
  Proporsi setiap kelas dalam data pengujian.
- Detection Rate
  Jumlah positif yang benar diprediksi dibagi dengan jumlah total positif sebenarnya untuk setiap kelas.
- Detection Prevalence
  Jumlah total prediksi positif dibagi dengan jumlah total data pengujian untuk setiap kelas.
- Balanced Accuracy
  Rata-rata dari sensitivitas dan spesifisitas. Ini memberikan ukuran keseluruhan kinerja model yang seimbang antara kedua kelas.

Analisis Hasil:

Hasil evaluasi model menunjukkan bahwa model Random Forest yang dilatih pada dataset tersebut memiliki kinerja yang buruk dalam memprediksi variabel dependen "Customer_rating". Akurasi yang rendah (20.63%) menunjukkan bahwa model memiliki tingkat kesalahan yang tinggi dalam melakukan prediksi dengan benar.

Selain itu, statistik kelas menunjukkan bahwa sensitivitas, spesifisitas, dan nilai prediksi positif dan negatif rendah untuk semua kelas. Hal ini menunjukkan bahwa model memiliki kesulitan dalam membedakan antara kelas-kelas yang ada dan memberikan prediksi yang akurat.

Nilai kappa yang sangat rendah (0.0079) menunjukkan tingkat kesepakatan yang sangat rendah antara prediksi model dan nilai sebenarnya.

Dalam keseluruhan, hasil evaluasi menunjukkan bahwa model Random Forest yang dilatih pada fitur-fitur yang terdaftar tidak memberikan prediksi yang baik untuk variabel dependen "Customer_rating". Kemungkinan terdapat faktor lain yang mempengaruhi variabel target yang tidak tercakup dalam fitur-fitur yang digunakan dalam model. Sebaiknya dilakukan analisis lebih lanjut untuk memahami dan memperbaiki performa model tersebut, seperti mempertimbangkan fitur-fitur tambahan, penyesuaian parameter model, atau menggunakan teknik pemrosesan data yang lebih canggih.

Dalam kasus ini, akurasi model sangat rendah, hanya sekitar 20.63%. Ada beberapa **faktor** yang dapat menyebabkan akurasi rendah dan beberapa cara untuk meningkatkannya. Berikut adalah beberapa faktor yang mungkin mempengaruhi akurasi rendah dan beberapa solusi yang dapat diterapkan sebagai berikut:

1. Kurangnya jumlah data pelatihan
   Jumlah data pelatihan yang terbatas dapat mengakibatkan model memiliki keterbatasan dalam mempelajari pola yang ada dalam data. Solusinya adalah mencoba mendapatkan lebih banyak data pelatihan jika memungkinkan. Jika tidak, teknik augmentasi data seperti flipping, rotasi, atau zoom dapat diterapkan untuk membuat variasi baru dari data yang ada.

2. Ketidakseimbangan kelas
   Jika jumlah sampel dalam setiap kelas tidak seimbang, model cenderung memiliki kecenderungan untuk memprediksi lebih banyak sampel ke kelas mayoritas. Hal ini dapat menyebabkan akurasi yang rendah karena performa model yang buruk dalam mengidentifikasi sampel pada kelas minoritas. Solusinya adalah menggunakan teknik penanganan ketidakseimbangan kelas seperti oversampling (duplikasi data minoritas) atau undersampling (mengurangi jumlah data mayoritas).

3. Parameter yang tidak dioptimalkan
   Hyperparameter pada model RandomForest mungkin tidak diatur dengan optimal. Penting untuk menjalankan eksperimen dengan berbagai nilai hyperparameter untuk menemukan konfigurasi yang

lebih baik. Beberapa hyperparameter yang relevan untuk RandomForest antara lain jumlah pohon (n_estimators), kedalaman pohon (max_depth), dan jumlah fitur yang dipertimbangkan saat mencari split (max_features).

4. Fitur yang tidak relevan atau kurang informatif
   Fitur yang tidak memiliki hubungan yang kuat dengan variabel target atau fitur yang memiliki banyak nilai yang hilang atau tidak relevan dapat mengurangi kinerja model. Penting untuk melakukan analisis fitur dan mempertimbangkan untuk menghilangkan atau mengganti fitur yang tidak memberikan kontribusi yang signifikan terhadap prediksi.

Kurangnya validasi silang: Dalam kasus ini, tidak ada informasi yang diberikan tentang metode validasi yang digunakan. Validasi silang seperti validasi silang lipat-k adalah penting untuk mengevaluasi model secara obyektif dan mengurangi kemungkinan overfitting. Pastikan untuk menggunakan metode validasi yang tepat dan mempertimbangkan validasi silang untuk mendapatkan estimasi kinerja yang lebih akurat.