Week12

Surenther

2024-11-16

Introduction

As outlined in the attached research paper (PDF), I am working to address research question using statistical methods in R. Below is the list of steps I plan to follow:

Import CSV

```r
# Import CSV
data <- read.table(file = "ObesityDataSet_raw_and_data_sinthetic.csv", header = TRUE, sep = ",")
```

What are the key eating habits that contribute to obesity?

Explanation

To determine the key eating habits that contribute to obesity, we can analyze the dataset by building a classification model where the target variable (NObeyesdad) represents different obesity levels. We will use logistic regression to identify which eating habits (e.g., FAVC for frequent consumption of high-calorie food, FCVC for frequency of vegetable consumption, NCP for the number of main meals, etc.) contribute the most to predicting obesity levels.Since NObeyesdad has multiple categories, we will use multinomial logistic regression.

```r
# Load necessary libraries
library(nnet,warn.conflicts = FALSE)       # For multinomial logistic regression
library(dplyr,warn.conflicts = FALSE)      # For data manipulation
library(caret,warn.conflicts = FALSE)      # For data partitioning and evaluation

# Convert categorical variables to factors
data$NObeyesdad <- as.factor(data$NObeyesdad)
data$Gender <- as.factor(data$Gender)
data$family_history_with_overweight <- as.factor(data$family_history_with_overweight)
data$FAVC <- as.factor(data$FAVC)
data$CAEC <- as.factor(data$CAEC)
data$SMOKE <- as.factor(data$SMOKE)
data$SCC <- as.factor(data$SCC)
data$CALC <- as.factor(data$CALC)
data$MTRANS <- as.factor(data$MTRANS)

# Select only the columns related to eating habits and the target variable
eating_habits <- data %>%
  select(NObeyesdad, FAVC, FCVC, NCP, CAEC, CH2O, SCC, FAF, TUE, CALC)
```

```r
# Split the data into training and testing sets
set.seed(123)
train_index <- createDataPartition(eating_habits$NObeyesdad, p = 0.8, list = FALSE)
train_data <- eating_habits[train_index, ]
test_data <- eating_habits[-train_index, ]
```

```r
# Fit the multinomial logistic regression model
multi_logit_model <- multinom(NObeyesdad ~ ., data = train_data)
```

```r
# Display the model summary to view the coefficients
summary(multi_logit_model)
```

```
## Call:
## multinom(formula = NObeyesdad ~ ., data = train_data)
## 
## Coefficients:
##                       (Intercept)    FAVCyes        FCVC        NCP
## Normal_Weight            6.617788 -0.6199175  -0.8348189 -0.3697195
## Obesity_Type_I           4.600854  1.4949138  -1.3088984 -0.8800900
## Obesity_Type_II          1.669404  1.8275387  -0.1652069 -0.2946703
## Obesity_Type_III      -600.478147  5.5214324 257.7828809  2.2601046
## Overweight_Level_I       3.753441  0.7167623  -0.7869032 -0.6588630
## Overweight_Level_II      4.938983 -0.8824850  -0.8796622 -0.6980641
##                       CAECFrequently      CAECno CAECSometimes       CH2O
## Normal_Weight              -4.010468  -2.7925880    -3.3007046 -0.2379815
## Obesity_Type_I             -6.779551  -4.3080855    -0.8134135  1.0750019
## Obesity_Type_II            -5.749618  -2.3854060    -0.1890964  0.1178594
## Obesity_Type_III            2.167541 -55.0580133     9.3256328  0.6668794
## Overweight_Level_I         -3.223557   0.4774209    -0.3167593  0.3026103
## Overweight_Level_II        -3.549379  -2.7999579    -0.3103387  0.7477121
##                             SCCyes         FAF        TUE CALCFrequently
## Normal_Weight           0.03775841 -0.04246758 -0.4627678       4.342755
## Obesity_Type_I         -2.47729928 -0.66884472 -0.7605506       3.860482
## Obesity_Type_II        -2.75976441 -0.54387747 -1.0711912       1.946846
## Obesity_Type_III      -11.14064658 -1.59471113 -0.1580952    -214.432996
## Overweight_Level_I      1.00422848 -0.24987689 -0.6087719       3.789574
## Overweight_Level_II    -1.84572592 -0.60268969 -0.5513724       4.057986
##                             CALCno CALCSometimes
## Normal_Weight            1.0889301    1.18610257
## Obesity_Type_I           0.6413624    0.09900974
## Obesity_Type_II         -0.4860548    0.20861366
## Obesity_Type_III      -195.4948262 -190.55032467
## Overweight_Level_I      -0.5404032    0.50427029
## Overweight_Level_II      0.6663469    0.21465016
## 
## Std. Errors:
##                       (Intercept)   FAVCyes      FCVC       NCP CAECFrequently
## Normal_Weight            1.017443 0.2569281 0.2016025 0.1334839       1.008146
## Obesity_Type_I           1.094078 0.4163707 0.2264772 0.1436808       1.521577
## Obesity_Type_II          1.209513 0.5065374 0.2312233 0.1544221       1.578156
## Obesity_Type_III         2.749965 1.3415469 9.9313880 0.7026885      33.803795
## Overweight_Level_I       1.122051 0.3413145 0.2211532 0.1430136       1.188411
## Overweight_Level_II    1.123225 0.2786957 0.2233784 0.1431896       1.203586
##                            CAECno CAECSometimes      CH2O     SCCyes       FAF
## Normal_Weight       1.303483e+00      1.000589 0.1707942  0.3609790 0.1258646
## Obesity_Type_I      1.685708e+00      1.087512 0.1918997  0.8169735 0.1370341
## Obesity_Type_II     1.775363e+00      1.209363 0.1916144  1.0507933 0.1425565
## Obesity_Type_III    1.499647e-11     33.779498 0.3052729 29.6180938 0.2466315
## Overweight_Level_I  1.393319e+00      1.150972 0.1886395  0.3718600 0.1390444
## Overweight_Level_II 1.683247e+00      1.164008 0.1872443  0.6071911 0.1370798
##                            TUE CALCFrequently    CALCno CALCSometimes
## Normal_Weight       0.1657138   8.420211e-01 0.4238714     0.4125993
## Obesity_Type_I      0.1720923   8.880493e-01 0.4583666     0.4561685
## Obesity_Type_II     0.1842109   1.027361e+00 0.5174834     0.5083800
## Obesity_Type_III    0.4227888   1.497575e-07 1.4713787     1.4458805
## Overweight_Level_I  0.1749263   8.784113e-01 0.4672265     0.4548510
## Overweight_Level_II 0.1720856   8.796771e-01 0.4598197     0.4602770
##
## Residual Deviance: 4357.733
## AIC: 4513.733
```

Findings

High-Calorie Food Consumption (FAVCyes): Strongly associated with higher obesity levels.
Vegetable Consumption (FCVC): Inversely related to obesity categories like Obesity_Type_I and Obesity_Type_II.
Water Consumption (CH2O): Minimal impact based on small coefficients.
Physical Activity (FAF): Reduces the likelihood of higher obesity levels.
Alcohol Consumption (CALC): Varies significantly, with strong positive or negative effects depending on the category.

```r
# Interpret the significance of each feature using the Z-values and p-values
z_values <- summary(multi_logit_model)$coefficients / summary(multi_logit_model)$standard.errors
p_values <- (1 - pnorm(abs(z_values), 0, 1)) * 2

# Combine coefficients, Z-values, and p-values for easier interpretation
coeff_summary <- data.frame(
  Feature = rownames(summary(multi_logit_model)$coefficients),
  Coefficient = as.vector(summary(multi_logit_model)$coefficients),
  Z_value = as.vector(z_values),
  P_value = as.vector(p_values)
)
```

```r
print(coeff_summary)
```

```
##                Feature   Coefficient       Z_value      P_value
## 1        Normal_Weight    6.61778801  6.504332e+00 7.803957e-11
## 2       Obesity_Type_I    4.60085440  4.205234e+00 2.608125e-05
## 3      Obesity_Type_II    1.66940447  1.380228e+00 1.675164e-01
## 4     Obesity_Type_III -600.47814668 -2.183585e+02 0.000000e+00
## 5   Overweight_Level_I    3.75344115  3.345160e+00 8.223498e-04
## 6  Overweight_Level_II    4.93898325  4.397146e+00 1.096834e-05
## 7        Normal_Weight   -0.61991751 -2.412805e+00 1.583027e-02
## 8       Obesity_Type_I    1.49491384  3.590343e+00 3.302426e-04
## 9      Obesity_Type_II    1.82753872  3.607905e+00 3.086799e-04
## 10    Obesity_Type_III    5.52143241  4.115721e+00 3.859716e-05
## 11  Overweight_Level_I    0.71676233  2.100006e+00 3.572834e-02
## 12 Overweight_Level_II   -0.88248499 -3.166483e+00 1.542945e-03
## 13       Normal_Weight   -0.83481893 -4.140916e+00 3.459210e-05
## 14      Obesity_Type_I   -1.30889842 -5.779382e+00 7.497561e-09
## 15     Obesity_Type_II   -0.16520692 -7.144907e-01 4.749238e-01
## 16    Obesity_Type_III  257.78288087  2.595638e+01 0.000000e+00
## 17  Overweight_Level_I   -0.78690323 -3.558182e+00 3.734310e-04
## 18 Overweight_Level_II   -0.87966224 -3.937991e+00 8.216671e-05
## 19       Normal_Weight   -0.36971953 -2.769768e+00 5.609616e-03
## 20      Obesity_Type_I   -0.88008995 -6.125315e+00 9.050420e-10
## 21     Obesity_Type_II   -0.29467030 -1.908214e+00 5.636362e-02
## 22    Obesity_Type_III    2.26010457  3.216368e+00 1.298243e-03
## 23  Overweight_Level_I   -0.65886300 -4.606997e+00 4.085264e-06
## 24 Overweight_Level_II   -0.69806409 -4.875102e+00 1.087521e-06
## 25       Normal_Weight   -4.01046827 -3.978064e+00 6.947860e-05
## 26      Obesity_Type_I   -6.77955098 -4.455607e+00 8.365623e-06
## 27     Obesity_Type_II   -5.74961839 -3.643251e+00 2.692160e-04
## 28    Obesity_Type_III    2.16754121  6.412124e-02 9.488737e-01
## 29  Overweight_Level_I   -3.22355728 -2.712495e+00 6.677886e-03
## 30 Overweight_Level_II   -3.54937938 -2.949003e+00 3.188008e-03
## 31       Normal_Weight   -2.79258802 -2.142404e+00 3.216096e-02
## 32      Obesity_Type_I   -4.30808548 -2.555653e+00 1.059888e-02
## 33     Obesity_Type_II   -2.38540605 -1.343616e+00 1.790726e-01
## 34    Obesity_Type_III  -55.05801328 -3.671399e+12 0.000000e+00
## 35  Overweight_Level_I    0.47742091  3.426500e-01 7.318618e-01
## 36 Overweight_Level_II   -2.79995789 -1.663427e+00 9.622708e-02
## 37       Normal_Weight   -3.30070456 -3.298762e+00 9.711236e-04
## 38      Obesity_Type_I   -0.81341353 -7.479584e-01 4.544852e-01
## 39     Obesity_Type_II   -0.18909644 -1.563604e-01 8.757489e-01
## 40    Obesity_Type_III    9.32563275  2.760738e-01 7.824914e-01
## 41  Overweight_Level_I   -0.31675927 -2.752103e-01 7.831547e-01
## 42 Overweight_Level_II   -0.31033866 -2.666120e-01 7.897679e-01
## 43       Normal_Weight   -0.23798150 -1.393382e+00 1.635043e-01
## 44      Obesity_Type_I    1.07500185  5.601896e+00 2.120200e-08
## 45     Obesity_Type_II    0.11785944  6.150865e-01 5.384976e-01
## 46    Obesity_Type_III    0.66687938  2.184535e+00 2.892296e-02
## 47  Overweight_Level_I    0.30261031  1.604172e+00 1.086761e-01
## 48 Overweight_Level_II    0.74771208  3.993243e+00 6.517577e-05
## 49       Normal_Weight    0.03775841  1.046000e-01 9.166932e-01
## 50      Obesity_Type_I   -2.47729928 -3.032288e+00 2.427073e-03
## 51     Obesity_Type_II   -2.75976441 -2.626363e+00 8.630281e-03
## 52    Obesity_Type_III  -11.14064658 -3.761433e-01 7.068104e-01
## 53  Overweight_Level_I    1.00422848  2.700555e+00 6.922388e-03
## 54 Overweight_Level_II   -1.84572592 -3.039778e+00 2.367530e-03
## 55       Normal_Weight   -0.04246758 -3.374070e-01 7.358101e-01
## 56      Obesity_Type_I   -0.66884472 -4.880862e+00 1.056230e-06
## 57     Obesity_Type_II   -0.54387747 -3.815172e+00 1.360883e-04
## 58    Obesity_Type_III   -1.59471113 -6.465967e+00 1.006528e-10
## 59  Overweight_Level_I   -0.24987689 -1.797101e+00 7.231956e-02
## 60 Overweight_Level_II   -0.60268969 -4.396634e+00 1.099424e-05
## 61       Normal_Weight   -0.46276781 -2.792572e+00 5.229081e-03
## 62      Obesity_Type_I   -0.76055062 -4.419434e+00 9.895994e-06
## 63     Obesity_Type_II   -1.07119118 -5.815026e+00 6.062450e-09
## 64    Obesity_Type_III   -0.15809516 -3.739341e-01 7.084534e-01
## 65  Overweight_Level_I   -0.60877186 -3.480162e+00 5.011105e-04
## 66 Overweight_Level_II   -0.55137236 -3.204058e+00 1.355050e-03
## 67       Normal_Weight    4.34275531  5.157538e+00 2.502186e-07
## 68      Obesity_Type_I    3.86048225  4.347148e+00 1.379189e-05
## 69     Obesity_Type_II    1.94684557  1.894996e+00 5.809297e-02
## 70    Obesity_Type_III -214.43299579 -1.431868e+09 0.000000e+00
## 71  Overweight_Level_I    3.78957405  4.314122e+00 1.602381e-05
## 72 Overweight_Level_II    4.05798619  4.613040e+00 3.968211e-06
## 73       Normal_Weight    1.08893012  2.569010e+00 1.019894e-02
## 74      Obesity_Type_I    0.64136241  1.399235e+00 1.617427e-01
## 75     Obesity_Type_II   -0.48605476 -9.392664e-01 3.475940e-01
## 76    Obesity_Type_III -195.49482622 -1.328651e+02 0.000000e+00
## 77  Overweight_Level_I   -0.54040319 -1.156619e+00 2.474279e-01
## 78 Overweight_Level_II    0.66634689  1.449148e+00 1.472962e-01
## 79       Normal_Weight    1.18610257  2.874708e+00 4.044007e-03
## 80      Obesity_Type_I    0.09900974  2.170464e-01 8.281722e-01
## 81     Obesity_Type_II    0.20861366  4.103499e-01 6.815493e-01
## 82    Obesity_Type_III -190.55032467 -1.317884e+02 0.000000e+00
## 83  Overweight_Level_I    0.50427029  1.108649e+00 2.675814e-01
## 84 Overweight_Level_II    0.21465016  4.663499e-01 6.409650e-01
```

Findings

Statistically Significant Features:

Features with small P-values (e.g., < 0.05) are statistically significant predictors of obesity levels.

For example:

FAVCyes (frequent consumption of high-calorie food) has a significant positive effect on several categories, such as Obesity_Type_I (Z = 3.59, P = 0.00033).

FCVC (frequency of vegetable consumption) has a significant negative effect for Obesity_Type_I (Z = -5.78, P = 7.50e-09).

Non-Significant Features:

Features with large P-values (> 0.05) are not significant predictors for certain categories.

For example:

CAECSometimes for Obesity_Type_II has a P-value of 0.875, indicating no significant effect.

Extreme Coefficients:

Some categories, like Obesity_Type_III, have extreme coefficients (e.g., -600 for the intercept or -195 for CALCno), which might indicate outliers or overfitting.

Effect Directions:

Positive coefficients: Features like CH2O (water consumption) positively influence categories like Obesity_Type_I and Overweight_Level_II. Negative coefficients: Features like FCVC (vegetable consumption) have a protective effect, reducing the likelihood of being in higher obesity categories.

```r
# Make predictions on the test data
predictions <- predict(multi_logit_model, newdata = test_data)

# Evaluate the accuracy of the model
conf_matrix <- table(Predicted = predictions, Actual = test_data$NObeyesdad)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
print(paste("Model accuracy on test data:", round(accuracy * 100, 2), "%"))
```

```
## [1] "Model accuracy on test data: 45.71 %"
```

```r
# Display the confusion matrix
print("Confusion Matrix:")
```

```
## [1] "Confusion Matrix:"
```

```r
print(conf_matrix)
```

```
##                       Actual
## Predicted             Insufficient_Weight Normal_Weight Obesity_Type_I
##   Insufficient_Weight                  18            13              3
##   Normal_Weight                         5             8              2
##   Obesity_Type_I                       11            12             46
##   Obesity_Type_II                       6             5             12
##   Obesity_Type_III                      6             8              4
##   Overweight_Level_I                    5             6              2
##   Overweight_Level_II                   3             5              1
##                       Actual
## Predicted             Obesity_Type_II Obesity_Type_III Overweight_Level_I
##   Insufficient_Weight               2                0                  2
##   Normal_Weight                     0                0                  2
##   Obesity_Type_I                   14                0                 20
##   Obesity_Type_II                  35                0                 19
##   Obesity_Type_III                  4               64                  2
##   Overweight_Level_I                2                0                 12
##   Overweight_Level_II               2                0                  1
##                       Actual
## Predicted             Overweight_Level_II
##   Insufficient_Weight                   3
##   Normal_Weight                         4
##   Obesity_Type_I                       25
##   Obesity_Type_II                      11
##   Obesity_Type_III                      0
##   Overweight_Level_I                    6
##   Overweight_Level_II                   9
```

Findings *Interpretation for Each Category:*
Insufficient_Weight:
Predicted correctly 18 times.
Misclassified as "Normal_Weight" 13 times, indicating confusion between these two classes.

Normal_Weight:
Predicted correctly 8 times.
Misclassified frequently as "Insufficient_Weight" (5) and "Obesity_Type_I" (12).

Obesity_Type_I:
Predicted correctly 46 times, but often confused with "Obesity_Type_II" (35) and "Normal_Weight" (12).

Obesity_Type_II:
Predicted correctly 12 times, but misclassified as "Obesity_Type_I" (14) and "Overweight_Level_I" (19).

Obesity_Type_III:
Predicted correctly 64 times, showing strong performance for this class.

Overweight_Level_I:
Predicted correctly 12 times but often confused with "Obesity_Type_I" (20).

Overweight_Level_II:
Predicted correctly only 9 times, with misclassifications spread across other categories.