Model 03 t-SNE and k-Means

2024-07-22

#Importing the necessary libraries

```
library(aricode)

## Warning: package 'aricode' was built under R version 4.3.3

library(mclust)

## Warning: package 'mclust' was built under R version 4.3.3

## Package 'mclust' version 6.1.1
## Type 'citation("mclust")' for citing this R package in publications.

library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.3.3

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.3

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

library(Rtsne)

## Warning: package 'Rtsne' was built under R version 4.3.3
```

#Reading the data sourced from Data Preparation file

```
data <- read.csv('modeldata.csv')
```

```
class_col<-ncol(data)

colnames(data)[class_col] <- "class"

data$class <- factor(data$class,
level=as.character(sort(unique(data$class))))

levels(data$class)

## [1] "0" "1"

head(data)

##   HouseHoldIncome Gender Age Race EthnicGroup Neighbourhood Degree
Hispanic
## 1               4      2  62    1           1             3      5
0
## 2               2      2  79    1           1             4      2
0
## 3               3      1  60    1           1             3      2
0
## 4               2      2  78    1           1             3      2
0
## 5               2      1  80    1           1             4      2
0
## 6               3      2  59    1           1             3      2
0
##   MaritalStatus JobStatus PhysicalHealth MentalHealth AttendChurchService
## 1             1         1              4            4                   3
## 2             5         0              4            4                   1
## 3             1         1              3            5                   5
## 4             1         0              3            3                   6
## 5             5         0              3            3                   5
## 6             1         1              4            4                   2
##   Bridge HealthDiscussions LiveAlone      BMI class
## 1      1                 1         0 29.63854     0
## 2      0                 1         0 33.77728     1
## 3      1                 1         0 71.40351     1
## 4      1                 1         0 26.17371     0
## 5      0                 1         1 24.82300     0
## 6      1                 1         0 28.48473     0
```

Splitting target variable in a different dataframe.

```
penddata <- dplyr::select(data, -class)
penclass <- data$class
```

Applying k-Means on the Original Dataset

```
set.seed(42)
kmeans_result <- kmeans(penddata, centers = 2, nstart = 25)
penddata <- penddata %>% mutate(Cluster = kmeans_result$cluster)
penddata$class <- data$class
head(penddata)

##   HouseHoldIncome Gender Age Race EthnicGroup Neighbourhood Degree
Hispanic
## 1               4      2  62    1           1             3      5
0
## 2               2      2  79    1           1             4      2
0
## 3               3      1  60    1           1             3      2
0
## 4               2      2  78    1           1             3      2
0
## 5               2      1  80    1           1             4      2
0
## 6               3      2  59    1           1             3      2
0
##   MaritalStatus JobStatus PhysicalHealth MentalHealth AttendChurchService
## 1             1         1              4            4                   3
## 2             5         0              4            4                   1
## 3             1         1              3            5                   5
## 4             1         0              3            3                   6
## 5             5         0              3            3                   5
## 6             1         1              4            4                   2
##   Bridge HealthDiscussions LiveAlone      BMI Cluster class
## 1      1                 1         0 29.63854       2     0
## 2      0                 1         0 33.77728       1     1
## 3      1                 1         0 71.40351       2     1
## 4      1                 1         0 26.17371       1     0
## 5      0                 1         1 24.82300       1     0
## 6      1                 1         0 28.48473       2     0
```

NMI Score: k-Means on original dataset

```
penddata$Cluster <- ifelse(penddata$Cluster == 2, 0, 1)
nmi_value <- NMI(as.factor(penddata$class), as.factor(penddata$Cluster))
cat("NMI:", nmi_value, "\n")

## NMI: 0.05257165
```

Appplying t-SNE on the Original Dataset

```
tsne_results <- Rtsne(penddata, perplexity = 50, check_duplicates = FALSE,
pca = TRUE, theta = 0.2)
```

```
tsne_df<-as.data.frame(tsne_results$Y)

tsne_df<- cbind(tsne_df, penclass)

means <- tsne_df %>%
  group_by(penclass) %>%
    summarise(mean_V1 = mean(V1),
              mean_V2 = mean(V2))
```

Applying k-Means on the t-SNE results

```
set.seed(123)
kmeans_result2 <- kmeans(tsne_df [,1:2], centers = 2, nstart = 25)
tsne_df <- tsne_df %>% mutate(Cluster = kmeans_result2$cluster)
head(tsne_df)

##           V1         V2 penclass Cluster
## 1  17.60348   7.610682        0       2
## 2 -14.99081 -21.187801        1       1
## 3  17.60528 -26.416937        1       2
## 4 -22.30187  -8.655713        0       1
## 5 -30.01260 -11.257196        0       1
## 6  22.71265   9.353244        0       2

tsne_df$Cluster <- ifelse(tsne_df$Cluster == 1, 0, 1)
nmi_value <- NMI(as.factor(tsne_df$penclass), as.factor(tsne_df$Cluster))
cat("NMI:", nmi_value, "\n")

## NMI: 0.03291776
```