Trabalho Prático 2: Clusterização

25/05/2025

Isabelle Fernandes de Oliveira
Lucas Dayrell de Andrade Machado

1. Contexto escolhido e sua relevância

O aumento global nos índices de obesidade tornou-se uma das principais preocupações de saúde pública nas últimas décadas. A obesidade está associada a diversas doenças crônicas, como diabetes tipo 2, hipertensão e problemas cardiovasculares. Com a crescente disponibilidade de dados relacionados a hábitos alimentares e estilos de vida, torna-se possível aplicar técnicas de mineração de dados para identificar padrões relevantes associados ao ganho de peso e perfis de risco.

O dataset Obesity Prediction, disponível no Kaggle *https://www.kaggle.com/datasets/ruchikakumbhar/obesity-prediction?resource=download*, reúne informações demográficas, comportamentais e nutricionais de indivíduos. Esse banco de dados oferece uma excelente oportunidade para aplicar técnicas de clusterização, permitindo a segmentação da população em grupos com características semelhantes, o que pode facilitar intervenções mais eficazes na prevenção e tratamento da obesidade.

2. Recursos disponíveis, requisitos, suposições, restrições, riscos e contingências

Recursos disponíveis:
- Dataset "Obesity Prediction" do Kaggle.
- Software R 4.4.2.
- Pacotes: stats, cluster, factoextra, fpc, NbClust, dbscan, mclust, tidyverse.
- RStudio: Ambiente integrado de desenvolvimento (IDE) para programação em R.
- Conhecimentos prévios sobre análise de dados e clusterização.

Requisitos:
- Realizar a preparação e tratamento dos dados.
- Aplicar e comparar algoritmos de clusterização (como K-means, DBSCAN,Hierarchical Clustering).
- Avaliar a coesão dos clusters utilizando métricas de avaliação dos agrupamentos.

Suposições:
- Os dados são representativos e confiáveis.
- As variáveis fornecidas possuem relação direta com a obesidade.

Restrições:
- O dataset é limitado a variáveis auto-relatadas, o que pode introduzir viés.
- O projeto será desenvolvido sem apoio de especialistas da área da saúde.

Riscos e contingências:
- Risco 1: a clusterização pode não gerar grupos claramente distintos.
- Contingência 1: testar diferentes algoritmos e métodos de pré-processamento.

- Risco 2: dados ausentes ou inconsistentes.
- Contingência 2: exclusão do registro

3. Objetivos da mineração de dados, detalhamento da tarefa e critérios de sucesso (Tarefa 1.3)

Objetivo:
- Segmentar indivíduos, por meio da aplicação de técnicas não supervisionadas (clusterização), com base em características relacionadas à obesidade, como hábitos alimentares, atividade física e dados demográficos, a fim de identificar perfis de risco distintos.
- Aplicar e comparar algoritmos de clusterização (como K-means, DBSCAN,Hierarchical Clustering).

Tarefa de Mineração de Dados:
1. Limpeza de dados.
2. Codificação de valores categóricos.
3. Padronização de variáveis numéricas.
4. Possível aplicação de Análise de Componentes Principais (PCA).
5. Utilização de algoritmos de agrupamento (K-means, DBSCAN, Hierarchical Clustering).
6. Avaliar a coesão dos clusters utilizando métricas de avaliação dos agrupamentos.
7. Comparar os resultados obtidos pelos diferentes algorítmos.

Critérios de sucesso:
- Clusters com alta separabilidade (baixa sobreposição entre grupos).
- Grupos que podem ser interpretados e relacionados ao risco de obesidade.

4. Descrição do projeto

O projeto consiste em aplicar técnicas de mineração de dados com foco em clusterização para analisar o dataset Obesity Prediction. O objetivo é agrupar os indivíduos com base em seus hábitos de vida e características demográficas, com o intuito de identificar perfis distintos de risco para a obesidade. Serão testados diferentes algoritmos de clusterização, e os resultados serão avaliados com base em métricas quantitativas e qualitativas.

5. Descrição e exploração dos dados

O dataset Obesity Prediction contém 2.111 linhas e 17 colunas. Cada linha representa o registro de indivíduos com as variáveis relacionadas a saúde e hábitos alimentares.

```
## Rows: 2,111
## Columns: 17
## $ Gender         <chr> "Female", "Female", "Male", "Male", "Male", "Male", "Fe~
## $ Age            <dbl> 21, 21, 23, 27, 22, 29, 23, 22, 24, 22, 26, 21, 22, 41,~
## $ Height         <dbl> 1.62, 1.52, 1.80, 1.80, 1.78, 1.62, 1.50, 1.64, 1.78, 1~
## $ Weight         <dbl> 64.0, 56.0, 77.0, 87.0, 89.8, 53.0, 55.0, 53.0, 64.0, 6~
## $ family_history <chr> "yes", "yes", "yes", "no", "no", "no", "yes", "no", "ye~
## $ FAVC           <chr> "no", "no", "no", "no", "no", "yes", "yes", "no", "yes"~
## $ FCVC           <dbl> 2, 3, 2, 3, 2, 2, 3, 2, 3, 2, 3, 2, 3, 2, 3, 3, 2, 2, 3~
## $ NCP            <dbl> 3, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 3, 1, 1, 4~
## $ CAEC           <chr> "Sometimes", "Sometimes", "Sometimes", "Sometimes", "So~
## $ SMOKE          <chr> "no", "yes", "no", "no", "no", "no", "no", "no", "no", ~
```

```
## $ CH2O           <dbl> 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, 2, 1, 2, 1, 2, 1~
## $ SCC            <chr> "no", "yes", "no", "no", "no", "no", "no", "no", "no", ~
## $ FAF            <dbl> 0, 3, 2, 2, 0, 0, 1, 3, 1, 1, 2, 2, 2, 2, 1, 2, 1, 0, 0~
## $ TUE            <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 2, 1, 0, 1, 1, 1, 0, 0, 0~
## $ CALC           <chr> "no", "Sometimes", "Frequently", "Frequently", "Sometim~
## $ MTRANS         <chr> "Public_Transportation", "Public_Transportation", "Publ~
## $ Obesity        <chr> "Normal_Weight", "Normal_Weight", "Normal_Weight", "Ove~
```

O que cada coluna representa está descrito abaixo:
- Gender: Sexo
- Age: Idade
- Height : Altura em metros
- Weight : peso em kg
- family_history : Algum membro da família sofreu ou sofre de sobrepeso?
- FAVC : Você come alimentos com alto teor calórico com frequência?
- FCVC : Você costuma comer vegetais em suas refeições?
- NCP : Quantas refeições principais você faz diariamente?
- CAEC : Você come algum alimento entre as refeições?
- SMOKE : Você fuma?
- CH2O : Quanta água você bebe diariamente?
- SCC : Você monitora as calorias que ingere diariamente?
- FAF: Com que frequência você pratica atividade física?
- TUE : Quanto tempo você usa dispositivos tecnológicos como celular, videogame, televisão, computador e outros?
- CALC : Com que frequência você consome bebidas alcoólicas?
- MTRANS : Qual meio de transporte você costuma usar?
- Obesity_level : Nível de obesidade

```r
qtde_registro_NA <- sum(apply(is.na(df), 1, any))
```

Quantidade de registros com alguma coluna NA: **0**.
Como pode ser visto, no dataset não há dados faltantes. Para todas as linhas, todas as colunas estão devidamente preenchidas.

O dataset possui variáveis categóricas e numéricas, sendo necessária a codificação adequada. Isso é feito no chunk abaixo com a transformação de dados character em factor.

```r
# trasnformacao de variavel character para categórica
for(i in 1:ncol(df)){
  if(is_character(df[,i]))
     df[,i] <- as.factor(df[,i])
}
```

Abaixo, segue a frequência dos registros para cada feature

```r
summary(df)
```

```
##     Gender          Age            Height          Weight       family_history
##  Female:1043   Min.   :14.00   Min.   :1.450   Min.   : 39.00   no : 385
##  Male  :1068   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47   yes:1726
##                Median :22.78   Median :1.700   Median : 83.00
##                Mean   :24.31   Mean   :1.702   Mean   : 86.59
##                3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
```

```
##                Max.   :61.00   Max.   :1.980   Max.   :173.00
##
##   FAVC           FCVC            NCP                CAEC      SMOKE
##  no : 245   Min.   :1.000   Min.   :1.000   Always    :  53   no :2067
##  yes:1866   1st Qu.:2.000   1st Qu.:2.659   Frequently: 242   yes:  44
##             Median :2.386   Median :3.000   no        :  51
##             Mean   :2.419   Mean   :2.686   Sometimes :1765
##             3rd Qu.:3.000   3rd Qu.:3.000
##             Max.   :3.000   Max.   :4.000
##
##       CH2O        SCC            FAF              TUE                CALC
##  Min.   :1.000   no :2015   Min.   :0.0000   Min.   :0.0000   Always    :   1
##  1st Qu.:1.585   yes:  96   1st Qu.:0.1245   1st Qu.:0.0000   Frequently:  70
##  Median :2.000              Median :1.0000   Median :0.6253   no        : 639
##  Mean   :2.008              Mean   :1.0103   Mean   :0.6579   Sometimes :1401
##  3rd Qu.:2.477              3rd Qu.:1.6667   3rd Qu.:1.0000
##  Max.   :3.000              Max.   :3.0000   Max.   :2.0000
##
##                    MTRANS                    Obesity
##  Automobile           : 457   Insufficient_Weight:272
##  Bike                 :   7   Normal_Weight      :287
##  Motorbike            :  11   Obesity_Type_I     :351
##  Public_Transportation:1580   Obesity_Type_II    :297
##  Walking              :  56   Obesity_Type_III   :324
##                               Overweight_Level_I :290
##                               Overweight_Level_II:290
```

A variável NObeyesdad pode ser usada para análise posterior, mas não será usada para guiar os clusters, já que o objetivo é a segmentação não supervisionada.

Será necessário normalizar variáveis numéricas como Age, Height, Weight, FAF, etc.

Análise exploratória por meio de histogramas, gráficos de dispersão e correlação será conduzida para entender a distribuição dos dados e possíveis relações entre as variáveis.