ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

CODEBOOK MÔ TẢ BỘ DỮ LIỆU DIABETES

Môn học: Thu thập và tiền xử lý dữ liệu - DS103.N21

Tên: Nguyễn Thị Huyền Trang

MSSV: 21520488

TP HỒ CHÍ MINH, 2023

MỤC LỤC

CODEBOOK

Bộ dữ liệu: Diabetes

Codebook:

Thông tin	**Nội dung**
Tên bộ dữ liệu	Diabetes patient records
Nguồn thu thập và cách thức thu thập	Từ 2 nguồn: 1. Máy đo tự động: có đồng hồ bấm giờ tự động, sẽ ghi lại chính xác thời gian tại lúc đo số liệu. 2. Thu thập bằng tay: Sử dụng bản ghi giấy (paper record), giờ được định sẵn vào các khung giờ: sáng (8:00), trưa (12:00), chiều (18:00) và tối (22:00)
Số thuộc tính	4
Thông tin tên các thuộc tính	Date: Ngày thu thập, định dạng: MM-DD-YYYY Time: Giờ thu thập, định dạng: XX:YY (24 giờ). Code: Mã code theo danh sách sau: 33 = Regular insulin dose 34 = NPH insulin dose 35 = UltraLente insulin dose 48 = Unspecified blood glucose measurement 57 = Unspecified blood glucose measurement 58 = Pre-breakfast blood glucose measurement 59 = Post-breakfast blood glucose measurement 60 = Pre-lunch blood glucose measurement 61 = Post-lunch blood glucose measurement 62 = Pre-supper blood glucose measurement 63 = Post-supper blood glucose measurement 64 = Pre-snack blood glucose measurement 65 = Hypoglycemic symptoms 66 = Typical meal ingestion 67 = More-than-usual meal ingestion 68 = Less-than-usual meal ingestion 69 = Typical exercise activity

	70 = More-than-usual exercise activity 71 = Less-than-usual exercise activity 72 = Unspecified special event Value: Gía trị thu thập được.
Thông tin tác giả	kahn@informatics.WUSTL.EDU (Internet) or 70333,34 (CompuServe)

Raw data:

Raw data gồm tập hợp các file: data-01, data-02, ... data-70.

Tidy data:

Tidy data sẽ được lưu lại thành file: diabets.csv.

Instruction list:

```
rm(list=ls())
myFiles <- list.files(path="diabetes-data", pattern="data")
data <- read.csv('diabetes-data/data-01', sep='\t',header = FALSE)
k = TRUE
# Tien hanh doc tung file
for (f in myFiles) {
  if (k==TRUE) {
    file <- read.csv(paste("diabetes-data/", f, sep=""), sep="\t",header = FALSE)
    k = FALSE
  }
  else {
    file <- rbind(file, read.csv(paste("diabetes-data/", f,
                                       sep=""), sep="\t", header = FALSE))
  }
}
dataset <- file
variables <- c("Date", "Time", "Code", "Value")
# Dat ten cho cot trong bo du lieu
colnames(dataset) <- variables
write.csv(dataset, file = "diabetes.csv")
```

THAM KHẢO

Diabetes Data Set, UCI Machine Learning Repository: Diabetes Data Set