ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

CODEBOOK MÔ TẢ BỘ DỮ LIỆU IRIS

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: Iris

Codebook:

Thông tin	**Nội dung**
Tên bộ dữ liệu	Iris Plants Database
Nguồn thu thập và cách thức thu thập	Tập dữ liệu chứa 3 lớp, mỗi lớp 50 cá thể, trong đó mỗi lớp đề cập đến loài cây Iris. Mỗi lớp phân tách tuyến tính với 2 lớp còn lại; lớp sau không phân tách tuyến tính với lớp trước.
Số thuộc tính	5
Kích thước	150
Thông tin tên các thuộc tính	Sepal Length: Chiều dài đài hoa Sepal Width: Chiều rộng đài hoa Petal Length: Chiều dài cánh hoa Petal Width: Chiều rộng cánh hoa Class: Phân lớp của hoa Iris
Thông tin tác giả	Nguồn: R.A. Fisher Người đóng góp: Michael Marshall (MARSHALL%PLU '@' io.arc.nasa.gov)

Raw data:

Raw data gồm file iris.data

Tidy data:

Xử lý file iris.data, dữ liệu sau khi xử lý được lưu dưới file iris.csv

Instruction list:

```
rm(list=ls())

myFiles <- list.files(path="iris-data/", pattern="iris.data")

file <- read.csv(paste("iris-data/", myFiles, sep=""), sep=",",header = FALSE)
dataset <- file
```

```
variables <- c("sepal length", "sepal width", "petal length", "petal width", "class")
s
colnames(dataset) <- variables
write.csv(dataset, file = "iris.csv",row.names = TRUE)
```

THAM KHẢO

Iris Plants Database, https://archive.ics.uci.edu/ml/machine-learning-databases/iris/