TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

KIỂM TRA THỰC HÀNH 02 – Đề 4

Môn: Thống kê máy tính và Ứng dụng

Ngày: 30/10/2024

Thời gian làm bài: 60 phút

Họ tên: ……Mã SV:……Số máy:……

Sinh viên nộp bài một file duy nhất ***MaSV_HoTen_TH02.ipynb***

Xem xét tập dữ liệu "Adult" được lưu trữ trên Kho lưu trữ máy học của UCI (https://archive.ics.uci.edu/ml/datasets/Adult), chứa khoảng 32.000 quan sát về các thông số tài chính khác nhau của dân số quốc tế. Chúng ta tập trung các cột sau:

age: Tuổi

sex: Giới tính (Male (nam) , Female (nữ))

country: Quốc gia (Mỹ ('United-States'))

income: mức thu nhập

1) ***(5đ)***
   a. Dữ liệu có bao nhiêu dòng, bao nhiêu cột. Tính tỷ lệ có thu nhập cao trên tổng số quan sát? Tính tỷ lệ ở Mỹ ('United-States') có thu nhập cao dựa trên tổng quan sát ở Mỹ, biết income >50K là có thu nhập cao? (1đ)
   b. Chọn ngẫu nhiên 2500 mẫu **cùng ở Mỹ**, sử dụng mẫu này để ước lượng tỷ lệ có thu nhập cao ở Mỹ với độ tin cậy là 90%? Để chọn mẫu ngẫu nhiên hãy sử dụng hàm *sample* với tham số *random_state=20*. (3đ)

      Gợi ý: **df[df['country'] == 'United-States'].sample(n, random_state=20)** để lấy ngẫu nhiên *n* phần tử từ *df* (*df* là DataFrame chứa dữ liệu)
   c. Từ kết quả ước lượng ở câu 1-b hãy cho nhận xét về ước lượng khoảng tỷ lệ có thu nhập cao ở Mỹ và giá trị thật tính từ dữ liệu. (1đ)
2) ***(5đ)*** Từ kết quả ước lượng khoảng tỷ lệ có thu nhập cao ở Mỹ, chúng ta có thể nhận xét rằng: "**Có hơn 23%** người lao động ở Mỹ có thu nhập cao" hay không, với mức ý nghĩa là 10%?

---------- Hết ----------

Các bài giống nhau bị điểm 0 Thực hành