Đề thi gồm 2 trang *- Đề thi được sử dụng tài liệu.*

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ THI GIỮA KỲ - ĐỀ 1

Môn thi : THỐNG KÊ MÁY TÍNH ỨNG DỤNG

Lớp/Lớp học phần: DHHTTT18ATT

Thời gian làm bài: **60 phút**

Họ và tên thí sinh ……; MSSV: ……; Số Máy: .....

Đề bài: DỮ LIỆU SINH VIÊN

Câu 1 (4 điểm) Trả lời các câu hỏi nội dung Phân Phối Xác Suất từ LMS

Câu 2 (6 điểm) Hệ thống giám sát yếu tố rủi ro hành vi (BRFSS - The Behavioral Risk Factor Surveillance System) là cuộc khảo sát qua điện thoại hàng năm đối với người trưởng thành ở Hoa Kỳ. BRFSS được thiết kế để xác định các yếu tố rủi ro trong dân số trưởng thành và báo cáo các xu hướng sức khỏe mới nổi.

Dữ liệu được cung cấp trong file *brfss_2000.csv* được khảo sát. Thông tin khảo sát trong bộ dữ liệu được ghi chú ở cuối tài liệu. Yêu cầu:

1. **(2 điểm)** Đọc dữ liệu. Hiển thị 5 dòng dữ liệu. Cho biết kích thước của dữ liệu. Với mỗi thuộc tính cho biết đó là thuộc tính định lượng hay thuộc tính phân loại; nếu là thuộc tính phân loại hãy cho biết thuộc tính có mấy loại. (sinh viên có thể trình bày bằng cách viết mã (code) hoặc ghi thông tin ở dạng văn bản)
2. **(1 điểm)** Khảo sát thuộc tính cân nặng (weight). Tìm các thông tin sau: trung bình, trung vị, miền giá trị (range), miền phân vị (IQR)
3. **(1 điểm)** Khảo sát thuộc tính hút thuốc (smoke100). Cho biết tỷ lệ người hút thuốc, tỷ lệ nam hút thuốc và nữ hút thuốc?
4. **(2 điểm)** Tạo thêm một cột có tên BMI biết rằng:

$$\text{BMI} = \frac{mass_{kg}}{height_{m}^{2}} = \frac{mass_{lb}}{height_{in}^{2}} \times 703$$

   a. Vẽ đồ thị histogram của BMI (với bins=50). Từ phân phối của BMI bạn hãy đưa ra nhận xét về các đối tượng khảo sát trong dữ liệu.
   b. Vẽ đồ thị boxplot về BMI theo thuộc tính tự đánh giá sức khỏe (genhlth). Dựa vào đồ thị đưa ra nhận xét.

Hướng dẫn:

- Tìm hiểu ý nghĩa của chỉ số BMI
- Có thể sử dụng seaborn để vẽ đồ thị boxplot

Lưu ý:

- Bài thi lưu theo định dạng ***<stt>_<hoten>_GK.ipynb***
- *Phần đầu file bài làm ghi thông tin: mã sinh viên, họ tên.*
- *Các câu trả lời nhận xét, bình luận gõ trực tiếp vào file bài làm*

---------- Hết ----------

brfss_2000

- **exerany:** 1 nếu người trả lời đã tập thể dục trong tháng qua và 0 nếu ngược lại.
- **hlthplan:** 1 nếu người trả lời có mua bảo hiểm y tế và 0 nếu ngược lại.
- **smoke100:** 1 nếu người trả lời đã hút ít nhất 100 điếu thuốc trong suốt cuộc đời và 0 nếu ngược lại.
- **height:** chiều cao của người trả lời tính bằng inch
- **weight:** trọng lượng của người trả lời tính bằng pound.
- **wtdesire:** trọng lượng mong muốn của người trả lời tính bằng pound.
- **age: tuổi**
- **gender: giới tính**
- **genhlth:** biểu thị sức khỏe tổng quát do người khảo sát tự đánh giá