Bài tập luyện tập

Bài tập 1: Bộ dữ liệu `housing.csv` cho biết thông tin về các khách hàng đến vay vốn mua nhà tại một công ty.
Bộ dữ liệu chứa thông tin với các cột như sau:

- `Loan_id` (Character): ID của khách hàng đến vay vốn.
- `Gender` (Character): Giới tính của khách hàng (Male/Female).
- `Married` (Character): Khách hàng đã kết hôn hay chưa (`Y/N`).
- `Dependents` (Character): Số người phụ thuộc (`0/1/2/3+`).
- `Education` (Character): Trình độ học vấn (Graduate/Under Graduate).
- `Self_Employed` (Character): Khách hàng có tự kinh doanh hay không (`Y/N`).
- `ApplicantIncome` (Numeric): Thu nhập của khách hàng.
- `CoapplicantIncome` (Numeric): Thu nhập của người cùng đứng tên khoản vay.
- `LoanAmount` (Numeric): Lượng tiền khách hàng muốn vay (Nghìn USD).
- `Loan_Amount_Term` (Numeric): Thời hạn trả khoản vay (Tháng).
- `Credit_History` (Numeric): Lịch sử tín dụng của khách hàng có đáp ứng yêu cầu hay không (`1/0`).
- `Property_Area` (Character): Khu vực nhà ở (Urban/Semi Urban/Rural).
- `Loan_status` (Character): Đơn vay vốn có được chấp nhận hay không (`Y/N`).

1. Hãy đọc dữ liệu vào trong R và đặt tên là `data`. Cho biết số hàng và số cột của bộ dữ liệu này.
2. Cột `Dependents` chứa các điểm dữ liệu có giá trị "`3+`". Hãy sửa các giá trị này thành "`3`".
3. Hãy cho biết cột nào trong `data` chứa giá trị `NA`. Đối với vector có kiểu dữ liệu ký tự, thay thế các giá trị `NA` bởi một giá trị bất kỳ khác `NA` trong vector đó. Đối với vector có kiểu dữ liệu số, thay thế các giá trị `NA` bởi giá trị trung bình của vector đó.
4. Thêm cột `Total_Income` là tổng thu nhập bằng tổng của 2 cột `ApplicantIncome` và `CoapplicantIncome` sau đó xóa đi 2 cột này.
5. So sánh số đơn đăng ký vay vốn được chấp thuận của khách hàng nam (Male) và khách hàng nữ (Female).
6. Để kiểm định xem giá trị trung bình của vector X có lớn hơn giá trị trung bình của vector Y hay không, người ta tính Test thống kê theo công thức sau:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

trong đó $\overline{X}$, s_X, n_X lần lượt là giá trị trung bình, độ lệch tiêu chuẩn và độ dài của vector X và tương tự với vector Y. Nếu $T > 1.96$ thì có thể nói rằng giá trị trung bình của X lớn hơn giá trị trung bình của Y và ngược lại.

Hãy kiểm định xem thu nhập trung bình (`Total_Income`) của nhóm khách hàng nam có lớn hơn của nhóm khách hàng nữ hay không.

7. Người ta muốn dự báo xem khoản vay của một khách hàng có được chấp nhận hay không (`Loan_status` là `Y` hay `N`) từ dữ liệu về `Credit_History`, `Total_Income` và `LoanAmount` của khách hàng đó. Sau khi áp dụng một mô hình dự báo, với từng khách hàng, ta có quy trình dự báo sau đây:

   - Bước 1: Thay thế giá trị của 2 biến `Total_Income` và `LoanAmount` bởi log của 2 biến này.
   - Bước 2: Tính

     $$h = -1.5453 + 3.2736 \times \text{Credit\_History} + 0.1086 \times \text{Total\_Income} - 0.2744 \times \text{LoanAmount}$$

     trong đó `Credit_History` = 1 nếu giá trị biến `Credit_History` là "`1`" và tương tự với giá trị "`0`".
   - Bước 3: Tính

     $$\phi(h) = \frac{1}{1 + e^{-h}}$$

   - Bước 4: So sánh $\phi(h)$ với 0.5. Nếu $\phi(h) \geq 0.5$ thì khoản vay được chấp nhận, hay dự báo cho `Loan_status` là `"Y"`, ngược lại là "`N`".

   Hãy thêm cột `Predict_Loan_status` là dự báo xem khoản vay của từng khách hàng có được chấp nhận hay không cho bộ dữ liệu `data`.

8. Để tính độ chính xác của dự báo, ta thực hiện phép tính sau:

   $$\text{accuracy} = \frac{TP + TN}{N}$$

   trong đó `accuracy` là độ chính xác của dự báo, `TP` là số dự báo là "`Y`" và thực tế là "`Y`", `TN` là số dự báo là "`N`" và thực tế là "`N`", `N` là tổng số dự báo. Hãy tính độ chính xác của dự báo ở câu i (Cột thực tế là `Loan_status`, cột dự báo là `Predict_Loan_status`).

Bài tập 2: Bộ dữ liệu `healthcare.csv` cho biết thông tin về các bệnh nhân đến khám chữa bệnh tại một bệnh viện.
Bộ dữ liệu chứa thông tin với các cột như sau:

- `ID` (Character): Mã bệnh nhân.
- `Age` (Numeric): Tuổi của bệnh nhân.
- `Gender` (Character): Giới tính của bệnh nhân (Male/Female).
- `Blood.Type` (Character): Nhóm máu của bệnh nhân (B-/B+/A-/A+/AB-/AB+/O-/0+).
- `Medical.Condition` (Character): Tình trạng bệnh lý của bệnh nhân (Cancer/Obesity/Asthma/Diabetes /Hypertension/Arthritis).

- `Date.of.Admission` (Character): Ngày bệnh nhân nhập viện theo định dạng "dd/mm/yyyy" (Ví dụ: 01/02/2020).
- `Insurance.Provider` (Character): Nhà cung cấp bảo hiểm y tế của bệnh nhân (Blue Cross/Medicare/Aetna/UnitedHealthcare/Cigna).
- `Billing.Amount` (Numeric): Viện phí - số tiền bệnh nhân phải thanh toán.
- `Admission.Type` (Character): Loại nhập viện (Urgent/Emergency/Elective).
- `Medication` (Character): Thuốc được kê đơn hoặc sử dụng trong quá trình điều trị (Aspirin/Ibuprofen/Penicillin/Paracetamol/Lipitor).
- `Test.Results` (Character): Kết quả xét nghiệm (Normal/Abnormal/Inconclusive).

1. Hãy đọc dữ liệu vào trong R và đặt tên là `data` sau đó loại bỏ các hàng chứa giá trị `NA`. Cho biết sau khi loại bỏ các giá trị `NA` thì bộ dữ liệu có bao nhiêu hàng và bao nhiêu cột.
2. Trích ra 2 bộ dữ liệu có giá trị ở biến `Gender` lần lượt là Male, Female và đặt tên tương ứng là `data.male`, `data.female`.
3. Làm tròn cột viện phí (`Billing.Amount`) đến một chữ số thập phân sau dấu phẩy và so sánh viện phí trung bình của nhóm bệnh nhân nam (Male) và nhóm bệnh nhân nữ (Female).
4. Cho biết số bệnh nhân ở từng nhóm máu.
5. Thêm cột `Classification.by.Age` phân nhóm các bệnh nhân theo tuổi. Các bệnh nhân từ 0 đến 14 tuổi được phân vào nhóm Pediatric (Trẻ em), các bệnh nhân từ 15 đến 47 tuổi được phân vào nhóm Young (Người trẻ), các bệnh nhân từ 48 đến 63 tuổi được phân vào nhóm Middle age (Trung niên), các bệnh nhân lớn hơn 64 tuổi được phân vào nhóm Elderly (Người già).
6. Viết một hàm nhận vào một tình trạng bệnh lý (`Medical.Condition`) và trả về độ tuổi trung bình của các bệnh nhân có bệnh lý đó. Áp dụng hàm vừa viết cho biết độ tuổi trung bình của các bệnh nhân mắc bệnh béo phì (Obesity), ung thư (Cancer) và tăng huyết áp (Hypertension).
7. Để kiểm định xem tỉ lệ bệnh nhân có kết quả xét nghiệm (`Test.Results`) là A (A có thể là Normal, Abnormal hoặc Inconclusive) có lớn hơn p_0 hay không, người ta tính Test thống kê theo công thức sau:

$$T = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0(1 - p_0)}}$$

trong đó $f = \frac{k}{n}$ với n là tổng số bệnh nhân và k là số bệnh nhân có kết quả xét nghiệm là A. Nếu $T > 1.645$ thì có cơ sở để kết luận rằng tỉ lệ bệnh nhân có kết quả xét nghiệm là A lớn hơn p_0 và ngược lại nếu $T \leq 1.645$ thì chưa có đủ cơ sở để kết luận điều này. Hãy viết 1 hàm nhận vào kết quả xét nghiệm A và tỉ lệ p_0 và kiểm định xem tỉ lệ bệnh nhân có kết quả xét nghiệm là A có lớn hơn p_0 hay không.

8. Áp dụng hàm đã viết ở câu k kiểm định xem tỉ lệ bệnh nhân có kết quả xét nghiệm là bình thường (Normal) có lớn hơn 50% hay không.