**Báo cáo tổng hợp dữ liệu PPG của Brno University of Technology Smartphone**

**I. Giới thiệu**

Cơ sở dữ liệu BUT PPG là một tập dữ liệu công khai được thiết kế để nghiên cứu về ước lượng nhịp tim từ tín hiệu quang thể tích đồ (PPG) và phân tích sóng mạch. Dữ liệu bao gồm các bản ghi đồng bộ của PPG, ECG và gia tốc kế (ACC) được thu thập từ nhiều đối tượng trong các điều kiện khác nhau. Đây là một vài thông tin tổng quan:

* **Tổng số người tham gia:** 50 đối tượng
* **Tổng số lần đo:** Mỗi người có nhiều lần đo khác nhau
* **Thiết bị ghi:**
  + Camera điện thoại thông minh cho tín hiệu PPG
  + Thiết bị ECG bên ngoài làm tín hiệu tham chiếu
* **Tần số lấy mẫu:**
  + **PPG:** 60 Hz (camera điện thoại)
  + **ECG:** 1 kHz (tín hiệu tham chiếu tiêu chuẩn vàng)
  + **ACC:** 100 Hz (được thêm vào trong các lần đo sau)

## **1. Cấu Trúc và Định Dạng Dữ Liệu**

Tập dữ liệu được tổ chức thành các thư mục, mỗi thư mục tương ứng với **một phiên đo cụ thể của một người tham gia**.

### **1.1 Quy Tắc Đặt Tên Tệp**

Mỗi thư mục được đặt tên theo mẫu:  
<Mã Đối Tượng><Số Lần Đo>

Ví dụ:

* 100001 → Đối tượng **100**, Lần đo **001**
* 102003 → Đối tượng **102**, Lần đo **003**

### **1.2 Các Tệp Trong Mỗi Thư Mục**

|  |  |
| --- | --- |
| **Tên Tệp** | **Mô Tả** |
| <ID>\_PPG.dat | Tín hiệu PPG thô (ghi từ camera điện thoại) |
| <ID>\_ECG.dat | Tín hiệu ECG tham chiếu (từ thiết bị ngoài) |
| <ID>\_ACC.dat | Dữ liệu gia tốc kế (theo dõi chuyển động) |
| <ID>.qrs | Vị trí phức bộ QRS trong ECG (dùng để phát hiện nhịp tim) |
| quality-hr-ann.csv | Nhãn chất lượng tín hiệu và nhịp tim tham chiếu |

## **1.3 Điều Kiện Đo Đạc**

Mỗi đối tượng được ghi lại trong **nhiều điều kiện khác nhau**, bao gồm:  
 Trạng thái nghỉ  
 Chuyển động có kiểm soát (tay, đầu, cơ thể, v.v.)  
 Nói chuyện trong khi đo  
 Điều kiện ánh sáng khác nhau (kiểm tra độ chính xác của PPG điện thoại)

## **1.4. Ứng Dụng Nghiên Cứu**

Cơ sở dữ liệu BUT PPG hữu ích cho các nghiên cứu về:  
 Ước lượng nhịp tim từ PPG  
 Loại bỏ nhiễu do chuyển động trong thiết bị đeo thông minh  
 Phân tích sóng mạch (pulse wave analysis)  
 Nhận dạng sinh trắc học và theo dõi sức khỏe qua PPG  
 Ứng dụng học máy trong xử lý tín hiệu y sinh

**II. Phân tích một số dữ liệu cơ bản**

Trong phần này, chúng ta sẽ thực hiện một phân tích tổng quát về dữ liệu trong bộ cơ sở dữ liệu PPG của Brno University of Technology (BUT PPG). Mục tiêu là hiểu rõ hơn về các đặc tính cơ bản của dữ liệu và rút ra những thông tin quan trọng từ các tín hiệu PPG, ECG, và gia tốc kế (ACC). Các phân tích sẽ tập trung vào các yếu tố sau:

1. **Phân phối độ tuổi của đối tượng tham gia**: Phân tích độ tuổi của những người tham gia sẽ giúp hiểu rõ hơn về độ tuổi và độ đa dạng của nhóm tham gia trong bộ dữ liệu này.
2. **Phân phối giới tính**: Phân tích sự phân bố giới tính sẽ cung cấp cái nhìn tổng quát về số lượng nam và nữ tham gia nghiên cứu.
3. **Chiều cao và cân nặng**: Sử dụng biểu đồ phân tán để khảo sát mối quan hệ giữa chiều cao và cân nặng của các đối tượng.
4. **Phân phối huyết áp**: Huyết áp là một yếu tố quan trọng trong việc nghiên cứu tín hiệu sinh lý học. Phân tích phân phối huyết áp sẽ giúp hiểu rõ hơn về các chỉ số huyết áp trong dữ liệu.
5. **Mức độ Glycaemia và SpO2**: Các yếu tố như mức đường huyết (glycaemia) và độ bão hòa oxy trong máu (SpO2) sẽ được phân tích để đánh giá tình trạng sức khỏe của các đối tượng.

#### **1. Phân Phối Độ Tuổi của Các Đối Tượng**

Một biểu đồ histogram đã được sử dụng để hiển thị sự phân phối độ tuổi của các đối tượng tham gia trong nghiên cứu. Biểu đồ này giúp xác định sự đa dạng về độ tuổi và kiểm tra xem liệu có sự tập trung vào một nhóm tuổi cụ thể nào hay không.

# Phân phối độ tuổi (Histogram)

plt.figure(figsize=(8, 5))

plt.hist(subject\_info\_clean['Age [years]'], bins=10, color='skyblue', edgecolor='black')

plt.title('Phân phối độ tuổi của các đối tượng tham gia')

plt.xlabel('Độ tuổi [năm]')

plt.ylabel('Tần suất')

plt.show()

#### **2. Phân Phối Giới Tính**

Biểu đồ cột (bar chart) được sử dụng để phân tích số lượng nam và nữ trong bộ dữ liệu. Biểu đồ này giúp nhận diện sự phân bố giới tính của các đối tượng tham gia.

# Phân phối giới tính (Bar chart)

plt.figure(figsize=(8, 5))

gender\_counts = subject\_info\_clean['Gender'].value\_counts()

gender\_counts.plot(kind='bar', color='lightgreen', edgecolor='black')

plt.title('Phân phối giới tính của các đối tượng tham gia')

plt.xlabel('Giới tính')

plt.ylabel('Tần suất')

plt.xticks(rotation=0)

plt.show()

#### **3. Mối Quan Hệ Giữa Chiều Cao và Cân Nặng**

Biểu đồ phân tán (scatter plot) được sử dụng để trực quan hóa mối quan hệ giữa chiều cao và cân nặng của các đối tượng. Điều này giúp nhận diện các mẫu trong dữ liệu và đánh giá sự phân bố của các đối tượng dựa trên các yếu tố cơ thể.

# Chiều cao và cân nặng (Scatter plot)

plt.figure(figsize=(8, 5))

plt.scatter(subject\_info\_clean['Height [cm]'], subject\_info\_clean['Weight [kg]'], color='purple', edgecolor='black', alpha=0.7)

plt.title('Mối quan hệ giữa chiều cao và cân nặng của các đối tượng')

plt.xlabel('Chiều cao [cm]')

plt.ylabel('Cân nặng [kg]')

plt.show()

#### **4. Phân Phối Huyết Áp**

Dữ liệu huyết áp được chia thành hai phần: huyết áp tâm thu (systolic) và huyết áp tâm trương (diastolic). Các biểu đồ histogram đã được sử dụng để phân tích sự phân bố của cả hai chỉ số này.

# Phân phối huyết áp (Histogram cho huyết áp tâm thu và tâm trương)

plt.figure(figsize=(8, 5))

plt.hist(subject\_info\_clean['Systolic BP'], bins=10, alpha=0.7, label='Huyết áp tâm thu', color='coral', edgecolor='black')

plt.hist(subject\_info\_clean['Diastolic BP'], bins=10, alpha=0.7, label='Huyết áp tâm trương', color='lightblue', edgecolor='black')

plt.title('Phân phối huyết áp (Tâm thu và Tâm trương)')

plt.xlabel('Huyết áp [mmHg]')

plt.ylabel('Tần suất')

plt.legend()

plt.show()

#### **5. Mức Đường Huyết và SpO2**

Các biểu đồ histogram cho mức glycaemia và SpO2 giúp phân tích sự phân bố của hai yếu tố này trong các đối tượng tham gia nghiên cứu.

# Mức Glycaemia (Histogram)

plt.figure(figsize=(8, 5))

plt.hist(subject\_info\_clean['Glycaemia [mmol/l]'], bins=10, color='orange', edgecolor='black')

plt.title('Phân phối mức đường huyết (Glycaemia)')

plt.xlabel('Mức đường huyết [mmol/l]')

plt.ylabel('Tần suất')

plt.show()

# Mức SpO2 (Histogram)

plt.figure(figsize=(8, 5))

plt.hist(subject\_info\_clean['SpO2 [%]'], bins=10, color='lightgreen', edgecolor='black')

plt.title('Phân phối mức độ bão hòa oxy trong máu (SpO2)')

plt.xlabel('SpO2 [%]')

plt.ylabel('Tần suất')

plt.show()

### **Kết Luận**

Phân tích cơ bản này giúp chúng ta có cái nhìn tổng quan về bộ dữ liệu PPG của Brno University of Technology, bao gồm sự phân phối độ tuổi, giới tính, mối quan hệ giữa chiều cao và cân nặng, cũng như các yếu tố sức khỏe khác như huyết áp, đường huyết và độ bão hòa oxy trong máu. Những thông tin này có thể giúp các nhà nghiên cứu và kỹ sư trong việc phát triển các mô hình ước lượng nhịp tim từ PPG, đồng thời cải thiện độ chính xác của các thiết bị theo dõi sức khỏe.

### **Ứng Dụng**

Bộ dữ liệu này rất hữu ích trong các nghiên cứu về ước lượng nhịp tim từ tín hiệu PPG, phân tích sóng mạch, loại bỏ nhiễu do chuyển động và phát triển các phương pháp học máy trong việc xử lý tín hiệu y sinh. Các biểu đồ và phân tích trên cung cấp một cái nhìn sâu sắc về các yếu tố sinh lý học quan trọng liên quan đến sức khỏe của con người.