|  |
| --- |
| Trường Đại học Khoa học tự nhiên – Khoa Công nghệ thông tin. |
| Đồ án thực hành cuối kỳ. |
| Programming for Data Science – Lập trình cho Khoa học dữ liệu. |

|  |
| --- |
| Nhóm 3TT  Tháng 12, 2024. |

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**📖👨‍💻📖**

**A blue and white logo

Description automatically generated**

**ĐỒ ÁN THỰC HÀNH CUỐI KÌ**

**Bộ môn:** Lập trình cho khoa học dữ liệu.

**Tên đề tài:**

*“Phân tích, đánh giá các yếu tố gây bệnh đau tim”.*

**Tên nhóm:** 3TT.

**Thành viên:**

1. 22120384 – Nguyễn Đình Trí.
2. 22120398 – Vũ Hoàng Nhật Trường.
3. 22120412 – Nguyễn Anh Tường.

# **Thông tin chung:**

1. **Bộ môn:** Lập trình cho khoa học dữ liệu.
2. **Giảng viên lý thuyết:** Thầy Phạm Trọng Nghĩa.
3. **Giảng viên thực hành:** Thầy Lê Nhựt Nam.
4. **Mã lớp:** 22\_21.
5. **Tên nhóm:** 3TT.
6. **Danh sách thành viên:**
   1. 22120384 – Nguyễn Đình Trí.
   2. 22120398 – Vũ Hoàng Nhật Trường.
   3. 22120412 – Nguyễn Anh Tường.
7. **Link github repository:** [“Click here to go to our github repository.”](https://github.com/richardnguyen0715/Pro4DS-FinalProject-HandleKaggleDatasetProblems)

Contents

[**ĐỒ ÁN THỰC HÀNH CUỐI KÌ** 2](#_Toc184925892)

[**Thông tin chung:** 3](#_Toc184925893)

[**Section 0:** *Bảng phân công công việc.* 5](#_Toc184925894)

[**I.** **Bảng phân công công việc:** 6](#_Toc184925895)

[**Section 1:** *Data Collection.* 7](#_Toc184925896)

[**I.** **Giới thiệu về tập dữ liệu:** 8](#_Toc184925897)

[**II.** **Một số thông tin khác:** 8](#_Toc184925898)

[**Section 2:** *Data Preprocessing.* 9](#_Toc184925899)

[**I.** **Ý nghĩa các cột trong tập dữ liệu:** 10](#_Toc184925900)

[**II.** **Xét các dòng trùng lặp:** 12](#_Toc184925901)

[**III.** **Kiểu dữ liệu của các cột trong tập dữ liệu:** 12](#_Toc184925902)

[**IV.** **Kiểm tra các giá trị thiếu khuyết của các cột:** 13](#_Toc184925903)

[**V.** **Các kết quả chỉ số liên quan đến thống kê:** 13](#_Toc184925904)

[**VI.** **Kiểm tra điều kiện tồn tại của các giá trị:** 14](#_Toc184925905)

[**VII.** **Xét các phân bố của các giá trị trong cột:** 15](#_Toc184925906)

[**Section 3:** *Data Exploration.* 18](#_Toc184925907)

# **Section 0:***A blue line drawing of a server Description automatically generated**Bảng phân công công việc.*

1. **Bảng phân công công việc:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Công việc** | **Người thực hiện** | **Tiến độ** | **Đóng góp** | **Note** |
| Data Collection | Nguyễn Anh Tường | 100% | 100% |  |
|  | | | | |
| Data Preprocessing | Nguyễn Anh Tường | 100% | 90% |  |
| Nguyễn Đình Trí | 10% |
|  | | | | |
| Data Exploration | Nguyễn Đình Trí | 100% | 50% |  |
| Hoàng Vũ Nhật Trường | 50% |

# **Section 1:** *Data Collection.*

***A blue line drawing of a server

Description automatically generated***

1. **Giới thiệu về tập dữ liệu:**
2. **Nguồn:** Kaggle.
3. **Lĩnh vực:** Y học – sức khỏe.
4. **Bản quyền:** CC0: Public Domain ( Được toàn quyền sử dụng tập dữ liệu. )
5. **Một số thông tin khác:**

**Chủ đề của tập dữ liệu là gì?**

Theo CDC, bệnh tim là một trong những nguyên nhân hàng đầu gây tử vong cho hầu hết các chủng tộc ở Hoa Kỳ (người Mỹ gốc Phi, người Mỹ bản địa, người Alaska bản địa và người da trắng). Khoảng một nửa người Mỹ (47%) có ít nhất 1 trong 3 yếu tố nguy cơ chính đối với bệnh tim: huyết áp cao, cholesterol cao, và hút thuốc lá. Các chỉ số quan trọng khác bao gồm tình trạng tiểu đường, béo phì (BMI cao), không vận động đủ, hoặc uống rượu quá nhiều. Việc xác định và ngăn chặn các yếu tố có tác động lớn nhất đến bệnh tim rất quan trọng trong chăm sóc sức khỏe. Bên cạnh đó, sự phát triển trong lĩnh vực tính toán cho phép áp dụng các phương pháp học máy để phát hiện "mẫu" trong dữ liệu, giúp dự đoán tình trạng của bệnh nhân.

**Nguồn gốc và xử lý của tập dữ liệu?**

Tập dữ liệu ban đầu đến từ CDC và là một phần quan trọng của Hệ thống Giám sát Yếu tố Nguy cơ Hành vi (BRFSS), thực hiện các cuộc khảo sát qua điện thoại hàng năm để thu thập dữ liệu về tình trạng sức khỏe của cư dân Hoa Kỳ. Theo CDC: "Được thành lập vào năm 1984 với 15 tiểu bang, BRFSS hiện thu thập dữ liệu ở tất cả 50 tiểu bang, Quận Columbia, và ba vùng lãnh thổ Hoa Kỳ. BRFSS hoàn thành hơn 400.000 cuộc phỏng vấn người lớn mỗi năm, khiến đây trở thành hệ thống khảo sát sức khỏe liên tục lớn nhất thế giới. Tập dữ liệu gần đây nhất bao gồm dữ liệu từ năm 2023. Trong tập dữ liệu này, tôi nhận thấy nhiều yếu tố (câu hỏi) ảnh hưởng trực tiếp hoặc gián tiếp đến bệnh tim, vì vậy tôi đã chọn lọc các biến liên quan nhất từ đó. Ngoài ra, tôi chia sẻ với bạn hai phiên bản của tập dữ liệu gần đây nhất: một phiên bản có NaN (dữ liệu thiếu) và một phiên bản không có NaN.

# **Section 2:** *Data Preprocessing.*

***A blue line drawing of a server

Description automatically generated***

1. **Ý nghĩa các cột trong tập dữ liệu:**

- **HeartDisease:** Biến mục tiêu chỉ ra tình trạng mắc bệnh tim của cá nhân. Giá trị:

- Yes: Có mắc bệnh tim.

- No: Không mắc bệnh tim.

- **BMI:** Chỉ số khối cơ thể (Body Mass Index), là thước đo dựa trên chiều cao và cân nặng để đánh giá tình trạng cơ thể (dư cân, bình thường, hoặc thiếu cân).

- **Smoking:** Thói quen hút thuốc lá.

- Yes: Có hút thuốc.

- No: Không hút thuốc.

- **AlcoholDrinking:** Thói quen uống rượu bia.

- Yes: Có uống rượu bia.

- No: Không uống rượu bia.

- **Stroke:** Tiền sử đột quỵ.

- Yes: Có từng bị đột quỵ.

- No: Không có tiền sử đột quỵ.

- **PhysicalHealth:** Số ngày trong 30 ngày gần nhất mà sức khỏe thể chất không tốt (đau ốm, mệt mỏi).

- **MentalHealth:** Số ngày trong 30 ngày gần nhất mà sức khỏe tinh thần không tốt (lo âu, trầm cảm, căng thẳng).

- **DiffWalking:** Khó khăn trong việc đi lại.

- Yes: Có khó khăn khi đi lại.

- No: Không có khó khăn khi đi lại.

- **Sex:** Giới tính của cá nhân.

- Male: Nam.

- Female: Nữ.

- **AgeCategory:** Nhóm tuổi của cá nhân (theo từng khoảng 5 năm, ví dụ: 40-44, 75-79).

- **Race:** Chủng tộc của cá nhân (ví dụ: White, Black).

- **Diabetic:** Tình trạng mắc bệnh tiểu đường.

- Yes: Có mắc bệnh tiểu đường.

- No: Không mắc bệnh tiểu đường.

- Yes (during pregnancy): Người bị tiểu đường trong thời kỳ mang thai, còn gọi là tiểu đường thai kỳ (gestational diabetes).

- No, Borderline diabetes:

- Người này có kết quả xét nghiệm đường huyết gần ranh giới chẩn đoán tiểu đường (tiền tiểu đường).

- Tiền tiểu đường là tình trạng đường huyết cao hơn mức bình thường nhưng chưa đủ để được coi là tiểu đường type 2.

- Đây là giai đoạn cảnh báo, nếu không thay đổi lối sống hoặc chế độ ăn, có thể tiến triển thành tiểu đường type 2.

- **PhysicalActivity:** Hoạt động thể chất thường xuyên.

- Yes: Có tham gia hoạt động thể chất.

- No: Không tham gia hoạt động thể chất.

- **GenHealth:** Đánh giá chung về sức khỏe của cá nhân.

- Excellent, Very good, Good, Fair, hoặc Poor.

- **SleepTime:** Số giờ ngủ trung bình mỗi ngày.

- **Asthma:** Tiền sử mắc bệnh hen suyễn.

- Yes: Có mắc bệnh hen suyễn.

- No: Không mắc bệnh hen suyễn.

- **KidneyDisease:** Tiền sử mắc bệnh thận.

- Yes: Có mắc bệnh thận.

- No: Không mắc bệnh thận.

- **SkinCancer:** Tiền sử mắc bệnh ung thư da.

- Yes: Có mắc bệnh ung thư da.

- No: Không mắc bệnh ung thư da.

1. **Xét các dòng trùng lặp:**

**A screenshot of a computer program

Description automatically generated**

Tập dữ liệu trước khi xử lý:



Sau khi thực hiện loại bỏ các dòng trùng lặp:



1. **Kiểu dữ liệu của các cột trong tập dữ liệu:**

A screen shot of a computer

Description automatically generated

1. **Kiểm tra các giá trị thiếu khuyết của các cột:**

A screenshot of a computer

Description automatically generated

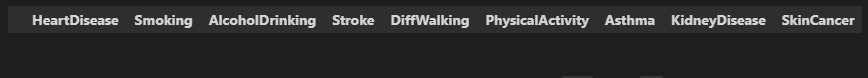
1. **Các kết quả chỉ số liên quan đến thống kê:**

A screenshot of a computer

Description automatically generated

1. **Kiểm tra điều kiện tồn tại của các giá trị:**

* Các cột như: `HeartDisease`, `Smoking`, `AlcoholDrinking`, `Stroke`, `DriffWalking`, `PhysicalActivity`, `Asthma`, `KidneyDisease`, `SkinCancer` có ý nghĩa là `có/không` ( có mắc hoặc không mắc hay có sử dụng hoặc không sử dụng,... ) vì thế các giá trị phải thuộc Yes/No. Ta cần kiểm tra xem trong các cột này có dòng nào tồn tại giá trị nằm ngoài giá trị `Yes` hoặc `No` hay không.



* **Kết luận**: không dòng nào vi phạm.
* Cột `Sex` thể hiện giới tính của bệnh nhân, giá trị có thể có là `Male` hoặc `Female`.

A screen shot of a computer

Description automatically generated

* **Kết luận:** chỉ có hai giá trị và đúng với thực tế.
* Các cột như `BMI`, `PhysicalHealth`, `MentalHealth`, `SleepTime` là những cột chỉ số xác định tình trạng sức khỏe của con người. Vì thế sẽ biến thiên tùy thuộc vào từng người. Nhưng điều kiện tiên quyết là giá trị của chúng phải lớn hơn hoặc bằng 0.



* **Kết luận:** không dòng nào vi phạm.

1. **Xét các phân bố của các giá trị trong cột:**

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

# **Section 3:** *Data Exploration.*

***A blue line drawing of a server

Description automatically generated***

4. Phân tích dữ liệu

Ở tập dữ liệu này, chúng ta sẽ cùng đi tìm hiểu về các yếu tố ảnh hưởng đến việc gây ra bệnh tim. Chúng ta cùng đi qua 3 góc nhìn dưới đây để hiểu rõ thêm và có được các thông tin cần thiết để từ đó rút ra và đúc kết kinh nghiệm cho bản thân trong việc phòng tránh bệnh về tim cũng như các bệnh lý đi kèm.

- Góc nhìn sức khỏe cá nhân

- Góc nhìn cộng đồng và yếu tố nhân khẩu học

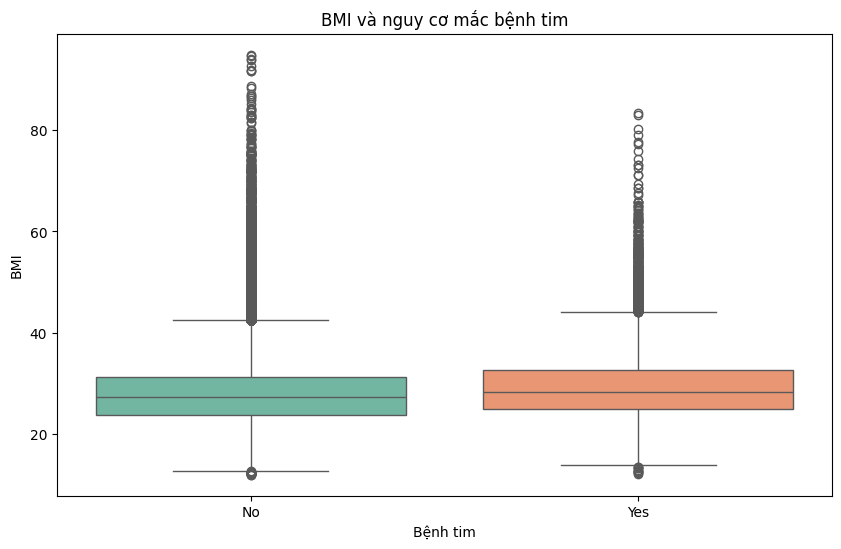
- Góc nhìn về các yếu tố hành vi và thói quen sống

4.1 Góc nhìn sức khỏe cá nhân

Mục tiêu: Hiểu rõ hơn về các yếu tố sức khỏe cá nhân và lối sống ảnh hưởng đến tình trạng sức khỏe, đặc biệt là bệnh tim.

4.1.1 Chỉ số BMI có liên quan đến tỉ lệ mắc bệnh tim không?

Chúng ta cùng quan sát biểu đồ boxplot sau đây:



***Nhận xét:***

* Phân bố BMI giữa hai nhóm (Yes/No)

- Nhóm "No" (không mắc bệnh tim):

- Giá trị trung vị (median) BMI của nhóm này nằm ở mức trung bình, khoảng từ 20- 30.

- Phân phối có sự hiện diện của nhiều outliers (các điểm ngoài phạm vi thông thường), đặc biệt ở phía trên (BMI > 40).

- Nhóm "Yes" (mắc bệnh tim):

- Trung vị BMI của nhóm này cũng nằm ở mức tương tự như nhóm "No" (khoảng 20- 30).

- Phân phối cũng có outliers tương tự, nhưng mức độ outliers (điểm trên cao) có vẻ ít hơn nhóm "No".

* Sự khác biệt giữa hai nhóm:

- Hai nhóm "Yes" và "No" có phân phối BMI tương đối giống nhau, với khoảng giá trị gần như tương đồng. Tuy nhiên:

- Nhóm "Yes" có xu hướng ít outliers hơn ở các mức BMI cao so với nhóm "No".

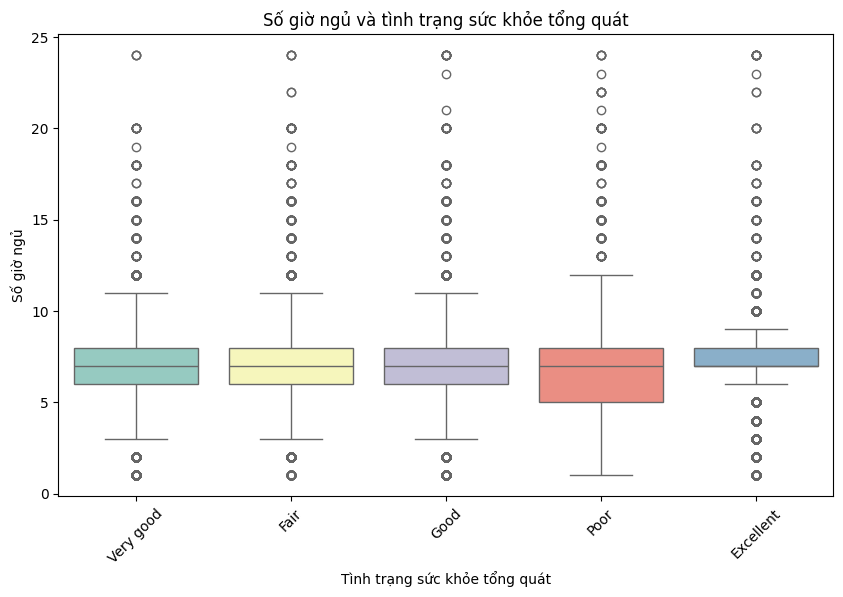
- Phạm vi BMI của hai nhóm khá tương đồng, điều này cho thấy BMI có thể không phải là yếu tố duy nhất ảnh hưởng đến nguy cơ mắc bệnh tim trong tập dữ liệu này.

* Khả năng liên hệ giữa BMI và nguy cơ mắc bệnh tim:

- Dựa trên biểu đồ, khó có thể khẳng định BMI có tác động lớn đến nguy cơ mắc bệnh tim, vì không có sự khác biệt rõ ràng giữa hai nhóm.

- Có thể cần kết hợp thêm các yếu tố khác (như huyết áp, cholesterol, lối sống, ...) để đánh giá toàn diện hơn về nguy cơ mắc bệnh tim.

4.2.2 Số giờ ngủ có ảnh hưởng đến tình trạng sức khỏe tổng quát như thế nào?



***Nhận xét:***

* Phân phối số giờ ngủ theo từng mức độ sức khỏe:

- Very good, Fair, Good, Poor, Excellent:

- Trung vị (median):

- Số giờ ngủ trung vị của các nhóm nằm trong khoảng từ 6 đến 8 giờ, đây là mức ngủ phổ biến.

- Độ trải rộng (IQR - Interquartile Range):

- Các nhóm có mức độ phân tán số giờ ngủ tương tự nhau, với hầu hết số giờ ngủ nằm trong khoảng 5-9 giờ.

- Outliers:

- Tất cả các nhóm đều có các outliers (những người ngủ ít hơn 4 giờ hoặc nhiều hơn 10 giờ), đặc biệt có một số trường hợp ngủ rất nhiều (hơn 20 giờ).

* Sự khác biệt giữa các mức độ sức khỏe:

- Nhóm "Poor" (sức khỏe kém):

- Nhóm này có độ trải rộng số giờ ngủ lớn hơn một chút, cho thấy những người trong nhóm này có thể có giấc ngủ không ổn định, cả ở mức ngủ ít và ngủ nhiều.

- Nhóm "Excellent" (sức khỏe rất tốt):

- Trung vị của số giờ ngủ ở nhóm này dường như thấp hơn một chút (gần 6 giờ), điều này có thể gợi ý rằng những người có sức khỏe rất tốt không nhất thiết ngủ nhiều, mà có thể do chất lượng giấc ngủ cao.

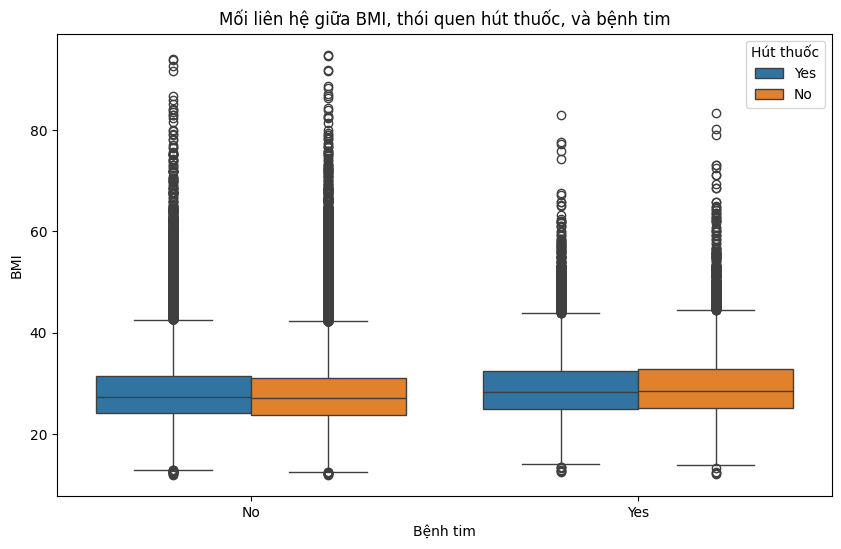
* Xu hướng liên quan giữa sức khỏe và số giờ ngủ:

- Không có sự khác biệt rõ ràng về số giờ ngủ giữa các nhóm sức khỏe tổng quát, nhưng có một vài quan sát:

- Sức khỏe tốt không chỉ phụ thuộc vào số giờ ngủ mà còn phụ thuộc vào chất lượng giấc ngủ hoặc các yếu tố khác như lối sống, chế độ ăn uống, và vận động.

- Nhóm "Poor" có xu hướng bao gồm nhiều người với số giờ ngủ nằm ở cả hai thái cực (ngủ quá ít hoặc quá nhiều), điều này có thể phản ánh tình trạng sức khỏe kém liên quan đến rối loạn giấc ngủ.

4.2.3 Mối liên hệ giữa BMI, thói quen hút thuốc và bệnh tim



* Phân phối BMI theo nhóm bệnh tim và thói quen hút thuốc:

- Nhóm không mắc bệnh tim ("No"):

- Người hút thuốc (Yes):

- Trung vị BMI của nhóm này nằm trong khoảng 20-30.

- Phân bố BMI của nhóm hút thuốc có vẻ tương tự với nhóm không hút thuốc, nhưng có thể xuất hiện một số giá trị outliers (BMI cao) ở phía trên.

- Người không hút thuốc (No):

- Trung vị BMI cũng tương tự, không khác biệt đáng kể so với nhóm hút thuốc.

- Số lượng outliers (BMI rất cao) ở nhóm này có vẻ hơi nhiều hơn so với nhóm hút thuốc.

- Nhóm mắc bệnh tim ("Yes"):

- Người hút thuốc (Yes):

- Trung vị BMI tương đương nhóm "No" (không mắc bệnh tim), nằm trong khoảng 20-30.

- Nhóm này có outliers ở phía BMI cao, nhưng không đáng kể hơn nhóm không hút thuốc.

- Người không hút thuốc (No):

- Trung vị BMI cũng tương tự nhóm hút thuốc, không có sự khác biệt lớn.

- Phân phối BMI không thay đổi nhiều so với nhóm không mắc bệnh tim.

* Sự khác biệt giữa các nhóm:

- BMI giữa nhóm "No" và "Yes" (bệnh tim):

- Không có sự khác biệt lớn về trung vị BMI giữa người mắc và không mắc bệnh tim. Nghĩa là BMI không phải là yếu tố duy nhất ảnh hưởng đến nguy cơ bệnh tim.

- Ảnh hưởng của hút thuốc:

- Sự khác biệt giữa nhóm hút thuốc và không hút thuốc là rất nhỏ trong cả hai nhóm bệnh tim và không bệnh tim. Điều này cho thấy, thói quen hút thuốc không tạo ra sự chênh lệch lớn về BMI.

* Outliers và phân phối:

- Outliers xuất hiện ở cả hai nhóm (mắc bệnh tim và không mắc bệnh tim), đặc biệt là ở các giá trị BMI rất cao (>40).

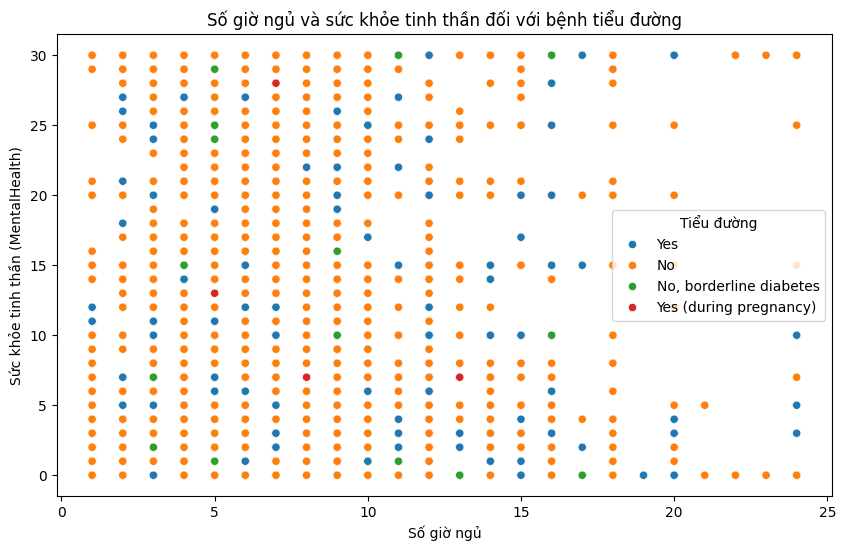
- Nhóm "Yes" (mắc bệnh tim) có vẻ có ít outliers hơn một chút so với nhóm "No", điều này có thể do số lượng mẫu hoặc đặc điểm của tập dữ liệu.

***Kết luận:***

- BMI: Không có sự khác biệt lớn về BMI giữa nhóm mắc và không mắc bệnh tim, cũng như giữa nhóm hút thuốc và không hút thuốc.

- Hút thuốc: Trong biểu đồ này, hút thuốc dường như không có tác động đáng kể đến chỉ số BMI hoặc sự khác biệt giữa các nhóm bệnh tim.

4.2.4 Số giờ ngủ và sức khỏe tinh thần có quan hệ gì đối với bệnh tiểu đường không?



* Phân bố số giờ ngủ và sức khỏe tinh thần:

- Số giờ ngủ:

- Phân bố trải dài từ 0 đến khoảng 24 giờ, nhưng tập trung chủ yếu trong khoảng 5-10 giờ, điều này phản ánh khoảng ngủ phổ biến trong dân số.

- Sức khỏe tinh thần:

- Giá trị sức khỏe tinh thần (MentalHealth) dao động từ 0 (tinh thần tốt) đến 30 (sức khỏe tinh thần rất kém). Phân phối rải đều, nhưng có nhiều điểm tập trung ở mức dưới 10.

* Mối quan hệ giữa sức khỏe tinh thần và bệnh tiểu đường:

- Nhóm "No" (không mắc tiểu đường):

- Đây là nhóm chiếm đa số (màu cam). Hầu hết phân bố của nhóm này tập trung trong khoảng 5-10 giờ ngủ và sức khỏe tinh thần khá tốt (MentalHealth < 10).oke

- Nhóm "Yes" (mắc tiểu đường):

- Nhóm này (màu xanh dương) xuất hiện ở nhiều khoảng sức khỏe tinh thần (từ 0 đến 30). Tuy nhiên, họ có vẻ phân bố tương tự về số giờ ngủ với nhóm "No".

- Nhóm "No, borderline diabetes" (tiền tiểu đường):

- Nhóm này (màu xanh lá) có phân bố tương đối ít và tập trung vào vùng sức khỏe tinh thần tốt hơn (thấp hơn 15), điều này có thể phản ánh mức độ ảnh hưởng thấp hơn so với người mắc bệnh tiểu đường hoàn toàn.

- Nhóm "Yes (during pregnancy)" (mắc tiểu đường trong thai kỳ):

- Nhóm này (màu đỏ) có số lượng rất ít, khó nhận xét chung. Tuy nhiên, một số điểm xuất hiện ở mức sức khỏe tinh thần cao (MentalHealth > 15).

4.3. Tác động tiềm năng:

- Số giờ ngủ và sức khỏe tinh thần:

- Có một số điểm bất thường, ví dụ: những người ngủ quá ít (0-2 giờ) hoặc quá nhiều (20-24 giờ) thường có sức khỏe tinh thần kém hơn (MentalHealth cao).

- Bệnh tiểu đường:

- Không có sự khác biệt rõ ràng về số giờ ngủ giữa các nhóm bệnh tiểu đường. Tuy nhiên, những người mắc tiểu đường hoặc tiền tiểu đường có xu hướng xuất hiện ở các giá trị sức khỏe tinh thần cao hơn (kém hơn).

- Tiểu đường trong thai kỳ (Yes during pregnancy) dường như không bị ảnh hưởng mạnh mẽ bởi số giờ ngủ, nhưng sức khỏe tinh thần có thể chịu ảnh hưởng.

4.4. Kết luận:

- Số giờ ngủ: Ngủ quá ít hoặc quá nhiều có thể liên quan đến sức khỏe tinh thần kém, bất kể trạng thái bệnh tiểu đường.

- Sức khỏe tinh thần: Những người mắc tiểu đường có vẻ dễ có sức khỏe tinh thần kém hơn so với người không mắc.