ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

DỰ ĐOÁN THỜI GIAN NGỦ TỪ LỐI SỐNG SỬ DỤNG CÁC MÔ HÌNH MÁY HỌC

Nhóm 24			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Hồ Thị Thanh Tuyền	22521627	KHMT
2	Hồ Thúy Nga	22520926	TTNT
3	Hoắc Công Minh	21522334	CNTT
4	Phạm Ngọc Ánh Hồng	22520492	TMĐT

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Mục tiêu chính của đồ án là xây dựng một mô hình dự đoán thời gian ngủ dựa trên các yếu tố liên quan đến lối sống và sức khỏe. Để thực hiện đồ án này, nhóm đã tiến hành thu thập dữ liệu, tiền xử lý, chọn lựa các đặc trưng quan trọng, và áp dụng các thuật toán học máy như Linear Regression, Random Forest, Gradient Boosting, Decision Tree Regression. Ngôn ngữ sử dụng trong đồ án là Python, kết hợp với các thư viện như Numpy, Pandas và Matplotlib để phân tích và trực quan hóa kết quả.

Bộ dữ liệu phân tích được tham khảo từ Kaggle. Tập dữ liệu được sử dụng là "Sleep Health and Lifestyle Dataset", bao gồm các thông tin cá nhân và thông tin sức khỏe của các đối tượng tham gia. Kết quả thực nghiệm cho thấy mô hình đạt độ chính xác cao trong việc dự đoán thời gian ngủ, đồng thời xác định được các yếu tố quan trọng ảnh hưởng đến thời gian ngủ.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu "Sleep Health and Lifestyle" chứa các biến liên quan đến thông tin cá nhân, giấc ngủ và thói quen sinh hoạt hàng ngày, cung cấp một cái nhìn tổng thể về mối quan hệ giữa thói quen, sức khỏe và chất lượng giấc ngủ của con người. Bộ dữ liệu phân tích được tham khảo tại Kaggle (https://www.kaggle.com/datasets)

Thống kê ban đầu của bộ dữ liệu gồm:

- Số dòng (số mẫu): 374
- Số cột (biến): 13

Trong đó:

- Biến phân loại (categorical variables): Gender, Occupation, BMI Category, Sleep Disorder.
- Biến số (numerical variables): Person ID, Age, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, Heart Rate, Daily Steps.

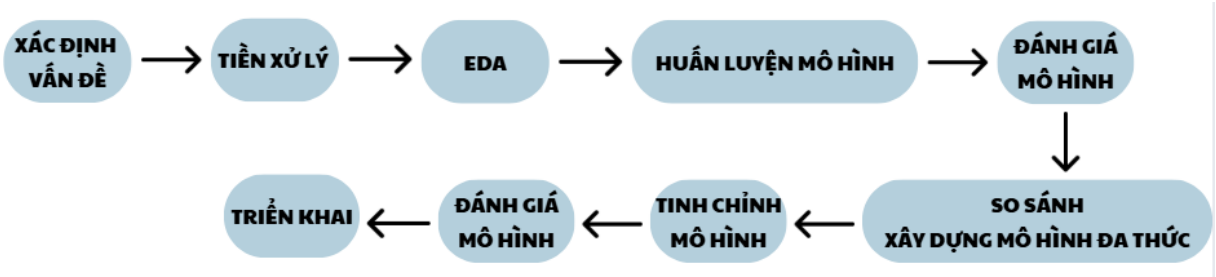
Mỗi mẫu dữ liệu là thông tin về sức khỏe giấc ngủ và lối sống của mỗi người và được mô tả thông qua các thuộc tính sau:

STT	Thuộc tính	Phạm vi	Mô tả
1	Person ID	Số nguyên từ 1 đến 374	Số thứ tự của từng người
2	Gender	Male, Female	Thông tin về giới tính
3	Age	Số nguyên từ 27 tới 59	Thông tin về độ tuổi

4	Occupation	Software Engineer, Doctor, Sales Representative, Teacher, Nurse, Engineer, Accountant, Scientist, Lawyer, Salesperson, Manager	Thông tin về nghề nghiệp
5	Sleep Duration	Số thập phân từ 5 đến 9	Thông tin về số giờ ngủ
6	Quality of Sleep	Số nguyên từ 1 tới 10	Đánh giá chất lượng giấc ngủ
7	Physical Activity Level	Số nguyên từ 30 đến 90	Số phút tham gia hoạt động thể chất hàng ngày
8	Stress Level	Số nguyên từ 1 đến 10.	Mức độ căng thẳng mà người đó cảm nhận
9	BMI Category	Overweight, Normal, Obese, Normal Weight	Phân loại chỉ số khối cơ thể
10	Blood Pressure	Số nguyên Huyết áp tâm thu: từ 115 - 145. Huyết áp tâm trương: từ 75-95.	Huyết áp của mỗi cá nhân, bao gồm: huyết áp tâm thu (số đầu tiên) và huyết áp tâm trương (số thứ hai)
11	Heart Rate	Số nguyên từ 65 đến 86	Nhịp tim khi nghỉ ngơi tính bằng nhịp mỗi phút
12	Daily Steps	Số nguyên từ 3000 đến 10000	Số bước đi mỗi ngày
13	Sleep Disorder	None, Sleep Apnea, Insomnia	Có mắc các chứng rối loạn giấc ngủ hay không

3. PHƯƠNG PHÁP PHÂN TÍCH

Nhóm đã tiếp cận việc phân tích dữ liệu thông qua một quy trình với các bước như sau:



Hình 1. Quy trình phân tích dữ liệu

3.1. Xác định vấn đề

Nhóm đã xác định vấn đề của đồ án là xây dựng một mô hình học máy có khả năng dự đoán thời gian ngủ của cá nhân dựa trên các đặc điểm về sức khỏe và lối sống.

3.2. Tiền xử lý dữ liệu

Trước khi tiến hành phân tích và xây dựng mô hình, nhóm đã thực hiện tiền xử lý bộ dữ liệu "Sleep Health and Lifestyle Dataset" để đảm bảo chất lượng và tính nhất quán. Các bước tiền xử lý cụ thể như sau:

- Xử lý giá trị khuyết: Các giá trị NaN trong cột "Sleep Disorder" được thay thế bằng chuỗi "None" nhằm thể hiện tình trạng "không có rối loạn giấc ngủ".
- Tiền xử lý cột "Blood Pressure": Cột "Blood Pressure" ban đầu chứa cả giá trị tâm thu và tâm trương, gây khó khăn trong việc phân tích riêng lẻ. Do đó nhóm chia cột này thành hai cột mới là "Systolic" (giá trị tâm thu) và "Diastolic" (giá trị tâm trương).
- Xóa cột không cần thiết: Cột "Person ID" chỉ đóng vai trò là một chỉ số duy nhất và không cung cấp thông tin hữu ích cho quá trình phân tích. Do đó, cột này đã được loại bỏ khỏi bộ dữ liệu.
- Chuẩn hóa giá trị: Trong cột "BMI Category", giá trị "Normal" và "Normal weight" có ý nghĩa tương đồng nhưng cách viết khác nhau sẽ được thống nhất về một giá trị duy nhất là "Normal".

3.3. Phân tích, thăm dò dữ liệu (EDA)

Đây là bước giúp nhóm hiểu rõ hơn về bộ dữ liệu và xác định mối quan hệ giữa các biến với nhau. Dưới đây là một số công việc mà nhóm đã thực hiện trong bước này:

- Thống kê mô tả cơ bản: mean, mode, median.
- Trực quan phân bố các biến định lượng và biến phân loại bằng biểu đồ histograms và boxplots.
- Vẽ biểu đồ tương quan (Correlation matrix) giữa các biến số.
- Phân tích sự tương quan giữa các biến số và biến phân loại đối với thời gian ngủ (Sleep Duration) bằng Pearson và ANOVA.
- Tìm hiểu sự ảnh hưởng của các biến với thời gian ngủ (Sleep Duration) và rút ra được các biến có ảnh hưởng đến thời gian ngủ nhiều nhất.

3.4. Lựa chọn và huấn luyện mô hình

Đối với đồ án này, quá trình lựa chọn và huấn luyện mô hình đóng vai trò quan trọng trong việc xây dựng một hệ thống dự đoán thời gian ngủ hiệu quả. Dưới đây là các bước chính:

- Chia tập dữ liệu thành tập huấn luyện (70%) và tập kiểm thử (30%).
- Chuẩn hóa các cột số bằng StandardScaler và mã hóa các cột phân loại thành dạng one-hot.
- Lựa chọn các mô hình để so sánh, bao gồm Linear Regression, Random Forest, Gradient Boosting, Decision Tree Regression.
- Xây dựng Pipeline với các mô hình đã lựa chọn ở trên.
- Huấn luyện Pipeline với tập huấn luyện.

3.5. Đánh giá và kiểm định mô hình

Các mô hình đã huấn luyện được đánh giá trên tập kiểm thử. Nhóm đã tiến hành các bước sau:

- Sử dụng Pipeline đã huấn luyện để dự đoán trên tập kiểm thử.
- Sử dụng kiểm định chéo (Cross-Validation) trên Pipeline đã xây dựng.
- Đánh giá mô hình dựa trên thang đo Mean Squared Error (MSE) và Rsquared (R2).
- Vẽ biểu đồ KDE (Kernel Density Estimation) nhằm so sánh phân phối của giá trị thực tế và giá trị dự đoán.

3.6. So sánh và xây dựng mô hình đa thức

Các mô hình sau khi được huấn luyện và đánh giá sẽ được so sánh để chọn ra mô hình phù hợp nhất. Nhóm đã tiến hành các bước sau:

- So sánh các mô hình gồm Linear Regression, Random Forest, Gradient Boosting, Decision Tree Regression dựa trên kết quả đạt được.
- Xây dựng mô hình đa thức từ bậc 1 đến bậc 7, chọn ra bậc đa thức tốt nhất.
- Nhóm thống nhất sử dụng mô hình Gradient Boosting với đa thức bậc 4 cho bài toán dự đoán thời gian ngủ.

3.7. Tinh chỉnh mô hình

Sau khi chọn mô hình và bậc đa thức, nhóm bắt đầu quá trình tinh chỉnh mô hình để cải thiện khả năng dự đoán và tổng quát hóa. Các cách thức nhóm đã sử dụng để tinh chỉnh mô hình như sau:

- Lựa chọn các tham số với GridSearchCV và RandomizedSearchCV.
- Áp dụng các tham số đã tinh chỉnh và so sánh.
- Xử lý các giá trị ngoại lệ.

3.8. Đánh giá mô hình sau tinh chỉnh

Sau khi thực hiện tinh chỉnh mô hình, nhóm tiến hành đánh giá lại mô hình sau tinh chỉnh. Các bước đánh giá như sau:

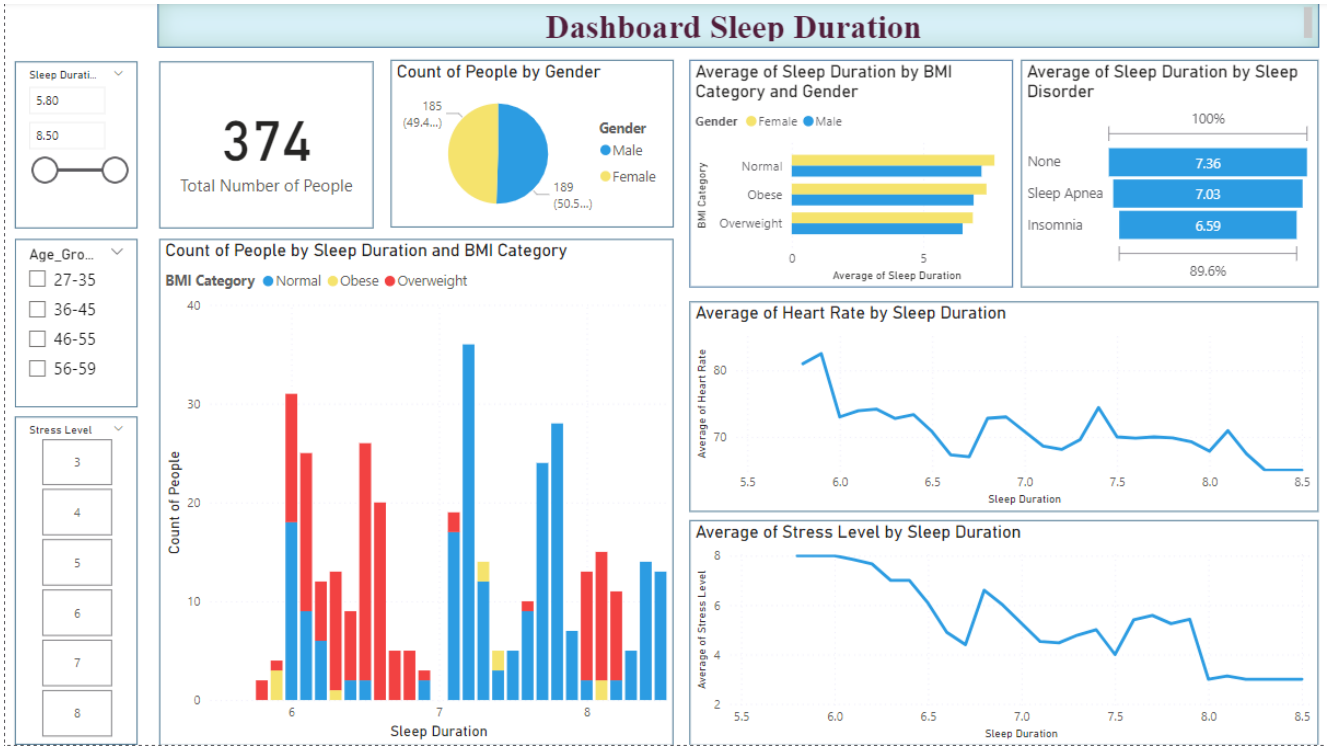
- Đánh giá mô hình dựa trên thang đo hiệu suất Mean Squared Error (MSE) và Rsquared (R2).
- Vẽ biểu đồ KDE (Kernel Density Estimation) nhằm so sánh phân phối của giá trị thực tế và giá trị dự đoán.

3.9. Triển khai mô hình

Sau khi mô hình đã được đánh giá kỹ lưỡng, nhóm triển khai mô hình để dự đoán thời gian ngủ trên bộ dữ liệu một lần nữa.

4. PHÂN TÍCH THĂM DÒ

Trong quá trình phân tích thăm dò, nhóm đã tập trung vào một số biến quan trọng để hiểu rõ hơn về tương quan và sự phân phối của chúng. Dưới đây là một số phát hiện chính và biến quan trọng đã được chọn lọc, có thể được thể hiện thông qua dashboard:



Hình 2. Dashboard

Nhóm sử dụng các loại biểu đồ để phân tích mối tương quan giữa các biến.

Card (Số lượng người tham gia): Hiển thị tổng số lượng người tham gia khảo sát (374 người). Thể hiện quy mô của tập dữ liệu và đánh giá mức độ tin cậy của kết quả phân tích.

Biểu đồ tròn: Thể hiện tỷ lệ phần trăm giới tính tham gia khảo sát. Đảm bảo dữ liệu đại diện cân bằng cho cả hai giới tính, không bị lệch về một nhóm cụ thể.

Biểu đồ thanh cụm: Hiển thị sự phân bố thời lượng ngủ (Sleep Duration) trung bình theo loại BMI và giới tính. So sánh sự khác biệt về thời lượng ngủ giữa các nhóm BMI và giới tính. Nữ có xu hướng ngủ nhiều hơn nam ở hầu hết các nhóm BMI. Người có chỉ số BMI bình thường có thời gian ngủ nhiều hơn so với người béo phì hoặc thừa cân.

Biểu đồ thanh ngang: Thể hiện thời gian ngủ trung bình theo từng loại rối loạn giấc ngủ. Đánh giá mức độ ảnh hưởng của rối loạn giấc ngủ đến thời lượng ngủ.

Biểu đồ đường: Thể hiện mối quan hệ giữa nhịp tim và thời lượng ngủ. Xác định xu hướng, nhịp tim có xu hướng giảm khi thời gian ngủ tăng lên.

Biểu đồ đường: Xác định xem mức độ căng thẳng cao có dẫn đến thời lượng ngủ ngắn hơn không.

Biểu đồ thanh: Hiển thị số lượng người tham gia khảo sát được phân loại theo thời gian ngủ và nhóm BMI. Giúp nhận biết nhóm BMI nào có sự phân bố thời lượng ngủ cao nhất.

Những phát hiện và phân tích thông qua quá trình phân tích thăm dò:

- Đối với các biến số: Mức độ căng thẳng ảnh hưởng lớn nhất đến thời lượng giấc ngủ. Mức độ căng thẳng tăng thì thời lượng giấc ngủ giảm. Nhịp tim có tác động yếu hơn, liên quan đến trạng thái thư giãn, chất lượng giấc ngủ. Chất lượng giấc ngủ góp phần kéo dài thời lượng ngủ.
- Đối với các biến phân loại: Người mắc những hội chứng khi ngủ có thời gian ngủ thấp hơn người bình thường. Nhóm BMI cũng ảnh hưởng, người bình thường có thời gian ngủ cao hơn.

5. KẾT QUẢ PHÂN TÍCH

Nhóm đã đặt mục tiêu là xây dựng một mô hình dự đoán thời gian ngủ đạt kết quả tốt dựa trên các yếu tố liên quan đến lối sống và sức khỏe. Để đạt được mục tiêu này, nhóm đã sử dụng các thư viện numpy, pandas, spicy, matplotlib, seaborn và sklearn để tiến hành các bước xử lý, trực quan hóa dữ liệu và lựa chọn mô hình phù hợp. Sau khi thực hiện so sánh các mô hình, nhóm lựa chọn sử dụng mô hình Gradient Boosting Regressor để phát triển cho bài toán và đạt được các kết quả cụ thể.

Kết quả đánh giá mô hình trên tập huấn luyện:

- MSE (Mean Squared Error): 0.00227
- R2 (R-squared): 0.99626

Kết quả đánh giá mô hình trên tập kiểm thử:

- MSE (Mean Squared Error): 0.00534

- R2 (R-squared): 0.99222

6. KẾT LUẬN

Nhóm đã tiến hành một quá trình phân tích dữ liệu chi tiết về các yếu tố liên quan đến lối sống và sức khỏe, sử dụng mô hình Gradient Boosting Regressor để dự đoán thời gian ngủ dựa trên nhiều yếu tố như độ tuổi, chất lượng giấc ngủ, mức độ căng thẳng, nhịp tim, giới tính, nghề nghiệp,... Kết quả của mô hình cho thấy độ chính xác tốt, với Mean Squared Error (MSE) đạt 0.00534 và R2 score đạt 0.99222.

Đồng thời, quá trình phân tích thăm dò đã rút ra một số phát hiện quan trọng về tương quan giữa các biến và mối quan hệ của chúng đối với thời gian ngủ. Các biến số quan trọng như chất lượng giấc ngủ, mức độ căng thẳng, nhịp tim được xác định là có ảnh hưởng lớn đến thời gian ngủ. Các biến phân loại như triệu chứng rối loạn giấc ngủ, chỉ số BMI, nghề nghiệp đều có tác động đáng kể đến thời gian ngủ.

TÀI LIỆU THAM KHẢO

[1] Scikit Learn. Link: https://scikit-learn.org/stable/supervised_learning.html .(15/11/2024).

[2] Matplotlib. Link: https://matplotlib.org/stable/index.html. (15/11/2024)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Hoắc Công Minh	- Phân chia công việc, lên kế hoạch thực hiện đồ án - Thu thập bộ dữ liệu - Tiền xử lý dữ liệu - Phân tích thăm dò bộ dữ liệu - Xây dựng mô hình - Đánh giá mô hình - Tinh chỉnh mô hình - Demo kết quả
2	Phạm Ngọc Ánh Hồng	- Tiền xử lý dữ liệu - Trực quan hóa dữ liệu - Phân tích thăm dò bộ dữ liệu - Xây dựng mô hình - Viết báo cáo
3	Hồ Thúy Nga	- Mô tả bộ dữ liệu - Tiền xử lý dữ liệu - Phân tích thăm dò bộ dữ liệu - Xây dựng mô hình - Đánh giá mô hình - Viết báo cáo
4	Hồ Thị Thanh Tuyền	- Thu thập bộ dữ liệu - Mô tả bộ dữ liệu - Tiền xử lý dữ liệu - Phân tích thăm dò bộ dữ liệu - Đánh giá mô hình - Làm slide