TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO KHAI PHÁ DỮ LIỆU
ĐỀ TÀI: ỨNG DỤNG CÂY QUYẾT ĐỊNH VÀ THUẬT TOÁN PHÂN CỤM TRONG ĐÁNH GIÁ TÁC ĐỘNG CỦA CÔNG NGHỆ ĐỐI VỚI TÂM LÝ CON NGƯỜI

Giảng viên hướng dẫn:	ThS. NGUYỄN THIỆN DƯƠNG
Sinh viên thực hiện:	
Dương Võ Anh Tài	6351071064
Nguyễn Trần Công Lý	6351071044
Nguyễn Đình Vương	6351071081
Nguyễn Quốc Anh	6351071003
Lớp:	Công nghệ thông tin
Khóa:	CQ.63.CNTT

TP.Hồ Chí Minh, năm 2025

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI TP. HỒ CHÍ MINH
BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO KHAI PHÁ DỮ LIỆU
ĐỀ TÀI: ỨNG DỤNG CÂY QUYẾT ĐỊNH VÀ THUẬT TOÁN PHÂN CỤM TRONG ĐÁNH GIÁ TÁC ĐỘNG CỦA CÔNG NGHỆ ĐỐI VỚI TÂM LÝ CON NGƯỜI

Giảng viên hướng dẫn:	ThS. NGUYỄN THIỆN DƯƠNG
Sinh viên thực hiện:	
Dương Võ Anh Tài	6351071064
Nguyễn Trần Công Lý	6351071044
Nguyễn Đình Vương	6351071081
Nguyễn Quốc Anh	6351071003
Lớp:	Công nghệ thông tin
Khóa:	CQ.63.CNTT

TP. Hồ Chí Minh, năm 2025

LỜI MỞ ĐẦU

Trong thời đại công nghệ số bùng nổ như hiện nay, các thiết bị điện tử và nền tảng trực tuyến đã trở thành một phần không thể tách rời trong đời sống con người. Việc sử dụng điện thoại, mạng xã hội, máy tính hay các ứng dụng giải trí mang lại nhiều tiện ích nhưng đồng thời cũng tiềm ẩn những tác động đáng kể đến sức khỏe tinh thần. Nhiều nghiên cứu chỉ ra rằng mức độ tiếp xúc với công nghệ có liên hệ chặt chẽ với sự gia tăng căng thẳng, lo âu, giảm chất lượng giấc ngủ và ảnh hưởng đến hành vi cảm xúc của người dùng.

Xuất phát từ vấn đề trên, đồ án "Ứng dụng cây quyết định và thuật toán phân cụm trong đánh giá tác động của công nghệ đối với tâm lý con người" được xây dựng nhằm sử dụng các phương pháp khai phá dữ liệu để phân tích, nhận dạng và dự đoán mức độ ảnh hưởng của việc sử dụng công nghệ đến trạng thái tâm lý. Bằng cách áp dụng mô hình Cây Quyết Định (Decision Tree) trong dự đoán mức độ căng thẳng, kết hợp với thuật toán Phân Cụm (K-Means) để phân loại các nhóm hành vi sử dụng công nghệ, đề tài hướng đến việc cung cấp một góc nhìn toàn diện về các yếu tố liên quan đến stress và cách chúng tương tác với thói quen sử dụng thiết bị số của con người.

Bộ dữ liệu gồm 5000 mẫu thông tin về thời gian sử dụng thiết bị, hoạt động giải trí, chất lượng giấc ngủ, thói quen sinh hoạt và các chỉ số sức khỏe tinh thần. Thông qua quy trình tiền xử lý dữ liệu, phân tích khám phá (EDA), xây dựng mô hình dự đoán và phân cụm, đồ án góp phần làm rõ mối quan hệ giữa hành vi công nghệ và trạng thái tâm lý, đồng thời gợi mở các hướng ứng dụng thực tiễn trong chăm sóc sức khỏe tinh thần và tối ưu hóa thói quen sử dụng công nghệ.

Đồ án không chỉ giúp củng cố kiến thức về khai phá dữ liệu và học máy, mà còn mang ý nghĩa thực tiễn trong bối cảnh con người ngày càng phụ thuộc vào công nghệ. Kết quả thu được kỳ vọng sẽ hỗ trợ người dùng, nhà nghiên cứu và các tổ chức trong việc xây dựng giải pháp cải thiện sức khỏe tinh thần và nâng cao chất lượng cuộc sống.

LỜI CẢM ƠN

Trong suốt quá trình thực hiện đồ án "Ứng dụng cây quyết định và thuật toán phân cụm trong đánh giá tác động của công nghệ đối với tâm lý con người", em đã nhận được rất nhiều sự hỗ trợ và hướng dẫn quý báu. Trước hết, em xin gửi lời tri ân sâu sắc đến thầy Dương, người đã tận tình định hướng, góp ý chuyên môn và tạo điều kiện thuận lợi để em có thể hoàn thành đề tài một cách tốt nhất. Sự nhiệt huyết, tận tâm và kiến thức chuyên sâu của thầy là nguồn động lực lớn giúp em vượt qua những khó khăn trong quá trình nghiên cứu và triển khai đồ án. Mặc dù đã cố gắng hết sức, đồ án khó tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý của thầy và quý thầy cô để em có thể hoàn thiện hơn trong tương lai.

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày …… tháng …… năm ……

Giảng viên hướng dẫn

Nguyễn Thiện Dương

MỤC LỤC

DANH MỤC VIẾT TẮT

STT	Viết tắt	Diễn giải	Ghi chú
1	CPU	Central Processing Unit	
2	GPU	Graphics processing unit	
3	TPU	Tensor Processing Unit	
4	PCA	Principal component analysis	
5			

DANH MỤC HÌNH ẢNH

CHƯƠNG 1: LÝ DO CHỌN DATASET VÀ GIỚI THIỆU TỔNG QUAN DATASET

1.1. Giới thiệu tổng quan Dataset

Bộ dữ liệu được sử dụng trong đồ án có tên "Tech Use and Stress Wellness Dataset", được xây dựng nhằm mô tả mối liên hệ giữa thói quen sử dụng công nghệ và các chỉ số sức khỏe tinh thần của người dùng. Dataset bao gồm 5000 mẫu dữ liệu với 25 thuộc tính, phản ánh nhiều khía cạnh trong hành vi sử dụng thiết bị số như thời gian dùng điện thoại, mạng xã hội, máy tính, chất lượng giấc ngủ, mức độ lo âu, trầm cảm và mức độ căng thẳng.

Các đặc trưng trong dataset có độ đa dạng cao, bao gồm cả biến định lượng và định tính, tạo điều kiện thuận lợi cho việc áp dụng những kỹ thuật khai phá dữ liệu như phân tích tương quan, phân cụm và mô hình hóa dự đoán. Đây là nguồn dữ liệu phù hợp để đánh giá tác động của công nghệ đối với tâm lý con người – đúng trọng tâm của đề tài đồ án.

1.1.1. Nguồn dữ liệu sử dụng

Bộ dữ liệu được thu thập và công bố bởi **Nagpal Prabhavalkar** trên nền tảng Kaggle – một kho dữ liệu lớn dành cho nghiên cứu khoa học dữ liệu và học máy. Người dùng có thể tải dataset hoàn toàn miễn phí và sử dụng cho mục đích học thuật, nghiên cứu hoặc thực hành.

Link dataset: https://www.kaggle.com/datasets/nagpalprabhavalkar/tech-use-and-stress-wellness

Dataset đi kèm mô tả chi tiết về từng trường dữ liệu, cho phép người thực hiện đồ án dễ dàng hiểu rõ ý nghĩa và phạm vi của từng thuộc tính liên quan đến hành vi công nghệ và sức khỏe tinh thần.

Hướng dẫn tải dataset thực hiện trong đồ án và các dataset khác của nhà cung cấp để tải dataset trên Kaggle, người dùng có thể thực hiện theo các bước sau:

Bước 1: Truy cập link dataset

Đi đến địa chỉ:

https://www.kaggle.com/datasets/nagpalprabhavalkar/tech-use-and-stress-wellness

Bước 2: Đăng nhập vào tài khoản Kaggle

Nếu chưa có tài khoản, người dùng cần tạo mới một tài khoản miễn phí để có quyền tải dữ liệu.

Bước 3: Nhấn "Download"

Ngay tại giao diện dataset, chọn nút **Download** để tải file dưới dạng .zip gồm:

- data.csv – dataset chính
- Các file mô tả và giấy phép sử dụng

Bước 4: Giải nén và đưa vào môi trường làm việc

Sau khi tải về, người dùng giải nén file và đặt dataset vào thư mục dự án (VD: data/data.csv) để sử dụng trong Python hoặc các công cụ phân tích khác.

Tải thêm dataset của cùng nhà cung cấp

Người dùng có thể xem các dataset khác của nhà cung cấp bằng cách:

1. Nhấn vào tên tác giả: **Nagpal Prabhavalkar** trên Kaggle
2. Truy cập mục *Datasets*
3. Lựa chọn và tải về các bộ dữ liệu tương tự phục vụ nghiên cứu nâng cao.

1.2 Mô tả chi tiết dữ liệu

Bộ dữ liệu được sử dụng trong nghiên cứu mang tên *Tech Use and Stress Wellness*, bao gồm 5000 quan sát với 25 thuộc tính phản ánh toàn diện hành vi sử dụng công nghệ và tình trạng sức khỏe tinh thần của người dùng. Các thuộc tính trong dataset được chia thành nhiều nhóm nội dung có quan hệ chặt chẽ với bài toán nghiên cứu.

Nhóm thuộc tính nhân khẩu học mô tả đặc điểm nền tảng của từng cá nhân, bao gồm mã định danh người dùng, độ tuổi, giới tính và loại khu vực sinh sống. Đây là những yếu tố cơ bản góp phần tạo nên sự đa dạng của tập dữ liệu và có thể ảnh hưởng gián tiếp đến thói quen sử dụng công nghệ.

Nhóm thuộc tính liên quan đến hành vi sử dụng công nghệ phản ánh mức độ tiếp xúc của người dùng với các thiết bị kỹ thuật số. Các trường dữ liệu như thời gian sử dụng thiết bị màn hình, thời lượng dùng điện thoại, máy tính xách tay, máy tính bảng,

tivi, cũng như thời gian dành cho mạng xã hội, công việc, giải trí và trò chơi điện tử cho phép đánh giá mức độ gắn bó của từng cá nhân với công nghệ trong đời sống hàng ngày.

Bên cạnh đó, dataset cũng bao gồm các thuộc tính liên quan đến lối sống và sức khỏe tinh thần, chẳng hạn như thời lượng ngủ, chất lượng giấc ngủ, mức độ hoạt động thể chất, số phút thực hành chánh niệm mỗi ngày và lượng caffeine tiêu thụ. Các yếu tố này có khả năng tác động trực tiếp tới tình trạng stress và thường được xem như những biến quan trọng trong các nghiên cứu về sức khỏe tinh thần. Ngoài ra, dữ liệu còn ghi nhận thông tin về việc người dùng có ăn uống lành mạnh hay sử dụng các ứng dụng chăm sóc sức khỏe hay không.

Đặc biệt, các chỉ số đánh giá tâm lý như mức độ lo âu và trầm cảm theo tuần, điểm đánh giá tâm trạng và điểm sức khỏe tinh thần tổng hợp đóng vai trò quan trọng trong việc mô tả trạng thái cảm xúc của người dùng. Biến mục tiêu của bài toán – stress_level – được biểu diễn dưới dạng thang điểm từ 1 đến 10, thể hiện mức độ căng thẳng của từng cá nhân tại thời điểm khảo sát.

Mô tả chi tiết từng cột dữ liệu

- Nhóm 1: Thông tin người dùng
  - user_id: Định danh duy nhất cho mỗi người dùng (từ 1 đến 5000).
  - age: Tuổi của người tham gia (trong dữ liệu thấy có từ trẻ em 15 tuổi đến người già 74 tuổi).
  - gender: Giới tính (Male, Female, Other).
  - location_type: Môi trường sống (Rural - Nông thôn, Suburban - Ngoại ô, Urban - Thành thị).
- Nhóm 2: Thời gian sử dụng thiết bị (Đơn vị: Giờ/Ngày)
  - daily_screen_time_hours: Tổng thời gian sử dụng màn hình trung bình mỗi ngày.
  - phone_usage_hours: Thời gian dùng điện thoại.
  - laptop_usage_hours: Thời gian dùng máy tính xách tay.
  - tablet_usage_hours: Thời gian dùng máy tính bảng.
  - tv_usage_hours: Thời gian xem tivi.
- Nhóm 3: Mục đích sử dụng thiết bị (Đơn vị: Giờ/Ngày)

- social_media_hours: Thời gian dùng mạng xã hội (Facebook, TikTok, v.v.).
- work_related_hours: Thời gian làm việc trên các thiết bị.
- entertainment_hours: Thời gian giải trí (xem phim, đọc tin tức).
- gaming_hours: Thời gian chơi điện tử.

- Nhóm 4: Giấc ngủ và Sức khỏe thể chất
  - sleep_duration_hours: Số giờ ngủ trung bình mỗi đêm.
  - sleep_quality: Chất lượng giấc ngủ (thang điểm 1-5, với 5 là rất tốt).
  - physical_activity_hours_per_week: Số giờ hoạt động thể chất (thể dục) mỗi tuần.
  - eats_healthy: Chế độ ăn uống lành mạnh (True/False).
  - caffeine_intake_mg_per_day: Lượng caffeine tiêu thụ mỗi ngày (mg).
- Nhóm 5: Chỉ số Sức khỏe tâm thần (Mental Health)
  - mood_rating: Tự đánh giá tâm trạng (thang điểm 1-10, cao là tốt).
  - stress_level: Mức độ căng thẳng (thang điểm 1-10, cao là rất căng thẳng).
  - mental_health_score: Điểm số sức khỏe tâm thần tổng quát (thường từ 0-100).
  - weekly_anxiety_score: Điểm lo âu hàng tuần.
  - weekly_depression_score: Điểm trầm cảm hàng tuần.
  - uses_wellness_apps: Có sử dụng các ứng dụng chăm sóc sức khỏe không (True/False).
  - mindfulness_minutes_per_day: Số phút thực hành chánh niệm/thiền mỗi ngày.

Nhìn chung, bộ dữ liệu có cấu trúc rõ ràng, phong phú về loại biến và phù hợp cho việc triển khai các kỹ thuật phân tích dữ liệu, mô hình hóa dự đoán và phân cụm người dùng nhằm nghiên cứu ảnh hưởng của công nghệ đối với sức khỏe tinh thần.

1.3 Mô tả mục đích bài toán

Mục đích chính của bài toán là phân tích và đánh giá tác động của việc sử dụng công nghệ đến mức độ căng thẳng của con người, đồng thời xây dựng các mô hình học máy nhằm dự đoán và phân loại người dùng theo đặc điểm hành vi công nghệ. Trong bối cảnh xã hội hiện đại, thời gian tiếp xúc với các thiết bị kỹ thuật số ngày càng tăng,

dẫn đến nhiều thay đổi trong thói quen sinh hoạt và sức khỏe tâm lý. Việc nghiên cứu các mối quan hệ này mang ý nghĩa quan trọng trong cả học thuật lẫn ứng dụng thực tiễn.

Trước tiên, bài toán hướng đến việc khám phá mối tương quan giữa hành vi sử dụng công nghệ và các chỉ số tâm lý bằng phương pháp phân tích dữ liệu khám phá. Thông qua quá trình này, nghiên cứu giúp làm rõ các yếu tố có khả năng ảnh hưởng mạnh đến mức độ stress, từ đó cung cấp nền tảng để phát triển các mô hình dự đoán đáng tin cậy.

Tiếp theo, bài toán tập trung xây dựng mô hình dự đoán mức độ căng thẳng của người dùng bằng thuật toán Cây Quyết Định. Thuật toán này không chỉ cho phép dự đoán stress_level dựa trên các dữ liệu đầu vào, mà còn mang lại khả năng diễn giải cao, giúp xác định các đặc trưng quan trọng và các điều kiện phân nhánh dẫn đến mức độ stress khác nhau. Điều này rất hữu ích trong việc hiểu sâu bản chất của vấn đề và đề xuất giải pháp giảm stress phù hợp.

Bài toán cũng bao gồm mục tiêu phân cụm người dùng dựa trên hành vi công nghệ bằng thuật toán K-Means. Cách tiếp cận này cho phép phân nhóm các đối tượng có đặc điểm sử dụng công nghệ tương đồng, từ đó hình thành các “hồ sơ hành vi" như nhóm sử dụng công nghệ nhiều cho công việc, nhóm sử dụng chủ yếu cho giải trí, hay nhóm có mức độ sử dụng cân bằng. Việc phân cụm giúp cung cấp cái nhìn toàn diện hơn về các mẫu hành vi và hỗ trợ đưa ra các đánh giá hoặc khuyến nghị mang tính cá nhân hóa.

Cuối cùng, bài toán đóng góp vào việc nhận diện nguy cơ stress tăng cao ở những nhóm người dùng cụ thể, hỗ trợ định hướng cho các giải pháp chăm sóc sức khỏe tinh thần trong môi trường số. Kết quả nghiên cứu không chỉ mang ý nghĩa học thuật trong lĩnh vực khai phá dữ liệu và học máy mà còn có tiềm năng ứng dụng trong thực tiễn, đặc biệt là trong việc xây dựng các chương trình hỗ trợ người dùng điều chỉnh thói quen công nghệ nhằm nâng cao chất lượng cuộc sống.

1.4 Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu đóng vai trò quan trọng trong việc đảm bảo chất lượng đầu vào cho các mô hình học máy. Bộ dữ liệu *Tech Use and Stress Wellness* tuy có cấu trúc rõ ràng và không chứa giá trị thiếu, nhưng vẫn cần trải qua nhiều bước xử lý nhằm

loại bỏ nhiễu, chuẩn hóa định dạng và chuẩn bị cho các thuật toán dự đoán và phân cụm. Phần này trình bày chi tiết các giai đoạn tiền xử lý dữ liệu được áp dụng trong đồ án.

1.4.1 Làm sạch dữ liệu

Bước đầu tiên của tiền xử lý dữ liệu tập trung vào việc kiểm tra và đảm bảo tính toàn vẹn của tập dữ liệu. Mặc dù bộ dữ liệu không xuất hiện giá trị bị khuyết, việc rà soát vẫn được tiến hành nhằm phát hiện các trường hợp dữ liệu bất thường như giá trị ngoại lai, sai lệch định dạng hoặc các quan sát không hợp lý.

Một số thuộc tính số học được kiểm tra bằng các kỹ thuật thống kê như quan sát phân bố, kiểm tra giá trị tối đa – tối thiểu và biểu đồ hộp nhằm phát hiện khả năng tồn tại của ngoại lệ. Trong trường hợp nhận diện được các giá trị bất thường có ảnh hưởng tiêu cực đến mô hình, các biện pháp xử lý như cắt ngưỡng (trimming) hoặc chuẩn hóa bằng phép biến đổi có thể được áp dụng.

Bên cạnh đó, dữ liệu định tính như *gender* và *location_type* được kiểm tra sự nhất quán về cách viết nhằm tránh phân tách sai các nhóm phân loại.

1.4.2 Tích hợp dữ liệu

Do dataset được cung cấp dưới dạng tệp đơn và có cấu trúc đồng nhất, bước tích hợp dữ liệu chủ yếu tập trung vào việc kết nối dữ liệu với các thông tin bổ sung cần thiết cho quá trình phân tích. Trong đồ án này, việc tích hợp dữ liệu không đòi hỏi ghép nối nhiều nguồn dữ liệu khác nhau, nhưng dữ liệu vẫn được xử lý theo dạng cấu trúc phù hợp với hệ thống mô hình hóa, chẳng hạn như gộp các biến liên quan vào cùng nhóm chức năng hoặc chuyển đổi định dạng dữ liệu để phục vụ thuật toán.

Quá trình tích hợp cũng bao gồm việc chuẩn bị các biến mới khi cần thiết. Chẳng hạn, nếu phục vụ mục tiêu trực quan hóa hoặc đánh giá mô hình, các thuộc tính được tạo thêm từ các biến gốc có thể được đưa vào phân tích, nhưng trong đồ án này dataset đã hoàn chỉnh nên không yêu cầu tạo thêm biến tổng hợp.

1.4.3 Thu giảm, rút gọn dữ liệu

Việc rút gọn dữ liệu nhằm giảm độ phức tạp trong quá trình xử lý và cải thiện hiệu suất của mô hình. Đối với bài toán hiện tại, quá trình thu giảm chủ yếu được thực hiện thông qua phân tích tương quan và loại bỏ những thuộc tính ít đóng góp hoặc mang tính trùng lặp cao.

Phân tích ma trận tương quan cho thấy các biến như mức độ lo âu theo tuần, mức độ trầm cảm theo tuần và thời gian sử dụng mạng xã hội có mức tương quan rất cao với biến mục tiêu *stress_level*. Tuy nhiên, thay vì loại bỏ các biến này, nghiên cứu giữ nguyên nhằm phục vụ việc phân tích mức độ ảnh hưởng của từng yếu tố đến stress. Ngược lại, các thuộc tính dạng chỉ số định danh như *user_id* được loại bỏ khỏi quá trình mô hình hóa vì không có giá trị dự báo.

Bên cạnh đó, dữ liệu định tính được chuyển đổi về dạng nhị phân hoặc dạng mã hóa số (encoding) khi cần, góp phần thu gọn không gian đặc trưng và chuẩn hóa cho thuật toán.

1.4.4 Dữ liệu sau tiền xử lý

Sau quá trình làm sạch, tích hợp và rút gọn dữ liệu, tập dữ liệu thu được có cấu trúc nhất quán và sẵn sàng cho các bước huấn luyện mô hình. Tập dữ liệu sau xử lý giữ nguyên các thuộc tính quan trọng như thông tin sử dụng công nghệ, chỉ số tâm lý và các yếu tố lối sống; đồng thời loại bỏ những yếu tố không mang tính đóng góp cho mô hình.

Các giá trị số được chuẩn hóa nhằm đảm bảo độ tương đồng về thang đo giữa các thuộc tính, đặc biệt quan trọng đối với thuật toán phân cụm K-Means vốn phụ thuộc mạnh vào khoảng cách Euclid. Các biến định tính được mã hóa bằng kỹ thuật one-hot hoặc mã hóa nhị phân để phù hợp với các thuật toán học máy truyền thống.

Kết quả là một tập dữ liệu có tính ổn định cao, cân bằng về cấu trúc và đáp ứng đầy đủ yêu cầu của các giai đoạn phân tích tiếp theo.

1.4.5 Chuẩn bị dữ liệu để huấn luyện và kiểm thử

Dữ liệu sau tiền xử lý được phân chia thành hai phần: tập huấn luyện và tập kiểm thử. Cách chia phổ biến là theo tỉ lệ 80 phần trăm dùng để huấn luyện mô hình và 20 phần trăm dùng để đánh giá khả năng tổng quát hóa của mô hình. Việc phân chia này giúp đảm bảo mô hình được đánh giá khách quan dựa trên dữ liệu chưa từng được sử dụng trong quá trình huấn luyện.

Sau khi phân chia, dữ liệu tiếp tục được chuẩn hóa bằng bộ chuẩn hóa (scaler) nhằm duy trì sự ổn định giữa các tập dữ liệu. Đối với mô hình Cây Quyết Định, chuẩn hóa không phải là yêu cầu bắt buộc, tuy nhiên bước này rất cần thiết cho thuật toán K-Means để đảm bảo phân cụm chính xác.

Cuối cùng, dữ liệu được chuyển vào các mô hình học máy tương ứng để tiến hành huấn luyện, kiểm thử và tối ưu hóa theo từng thuật toán.

1.5 Mô tả chi tiết các thuộc tính trong dataset

Bộ dữ liệu *Tech Use and Stress Wellness* bao gồm 25 thuộc tính phản ánh nhiều khía cạnh khác nhau của người dùng, từ đặc điểm nhân khẩu học, hành vi sử dụng công nghệ, thói quen cuộc sống đến các chỉ số tâm lý. Các thuộc tính này được thiết kế nhằm hỗ trợ việc nghiên cứu mức độ ảnh hưởng của công nghệ đến sức khỏe tinh thần và đặc biệt là mức độ căng thẳng. Mỗi thuộc tính mang một vai trò riêng trong việc cung cấp thông tin đầu vào cho quá trình phân tích và mô hình hóa.

Trước hết, nhóm thuộc tính nhân khẩu học gồm *user_id*, *age*, *gender* và *location_type* đóng vai trò mô tả đặc điểm nền của từng cá nhân trong bộ dữ liệu. Thuộc tính *user_id* chỉ được dùng để phân biệt các mẫu dữ liệu và không mang ý nghĩa dự báo. Thuộc tính *age* thể hiện độ tuổi của người dùng và có thể ảnh hưởng đến mức độ sử dụng công nghệ cũng như tình trạng sức khỏe tinh thần. Thuộc tính *gender* phản ánh giới tính, trong khi *location_type* mô tả môi trường sống, phân biệt giữa khu vực thành thị, ngoại ô và nông thôn. Các yếu tố nhân khẩu học này giúp tăng mức độ đa dạng và đại diện của dữ liệu.

Tiếp theo, nhóm thuộc tính liên quan đến hành vi sử dụng công nghệ phản ánh mức độ tiếp xúc với các thiết bị kỹ thuật số. Thuộc tính *daily_screen_time_hours* cho biết tổng thời gian người dùng sử dụng thiết bị màn hình trong ngày. Các thuộc tính như *phone_usage_hours*, *laptop_usage_hours*, *tablet_usage_hours* và *tv_usage_hours* mô tả thời lượng sử dụng từng loại thiết bị cụ thể. Ngoài ra, các biến *social_media_hours*, *work_related_hours*, *entertainment_hours* và *gaming_hours* thể hiện mức độ tham gia của người dùng vào các hoạt động kỹ thuật số mang tính chất xã hội, công việc, giải trí hoặc trò chơi điện tử. Đây là nhóm thuộc tính quan trọng giúp nghiên cứu trực tiếp mức độ gắn bó của cá nhân với công nghệ, từ đó định lượng tác động của chúng đến tình trạng căng thẳng.

Bên cạnh hành vi công nghệ, bộ dữ liệu còn bao gồm các thuộc tính mô tả lối sống và sức khỏe thể chất của người dùng. Thuộc tính *sleep_duration_hours* thể hiện thời gian ngủ trung bình mỗi ngày và có mối liên hệ mật thiết với sức khỏe tinh thần. Thuộc tính *sleep_quality* đánh giá chất lượng giấc ngủ theo thang điểm từ 1 đến 10,

trong khi *physical_activity_hours_per_week* phản ánh mức độ vận động của người dùng trong tuần. Thuộc tính *mindfulness_minutes_per_day* thể hiện số phút thực hành chánh niệm, một yếu tố được chứng minh có khả năng giảm mức độ căng thẳng. Bên cạnh đó, lượng caffeine tiêu thụ hằng ngày được thể hiện qua thuộc tính *caffeine_intake_mg_per_day*, và hai thuộc tính *eats_healthy* cùng *uses_wellness_apps* cho biết thói quen ăn uống lành mạnh và việc sử dụng ứng dụng chăm sóc sức khỏe. Các thuộc tính này giúp mô hình đánh giá toàn diện hơn về lối sống của người dùng, một yếu tố có ảnh hưởng đáng kể đến stress.

Bộ dữ liệu cũng bao gồm nhóm thuộc tính đánh giá trực tiếp tình trạng tâm lý, bao gồm *weekly_anxiety_score*, *weekly_depression_score* và *mood_rating*. Đây là những chỉ số quan trọng phản ánh mức độ lo âu, trầm cảm và trạng thái cảm xúc trong cuộc sống hằng ngày. Thuộc tính *mental_health_score* tổng hợp một số yếu tố liên quan đến sức khỏe tinh thần, giúp cung cấp thêm góc nhìn về điều kiện tâm lý chung của người dùng.

Cuối cùng, *stress_level* chính là thuộc tính mục tiêu của bài toán. Thuộc tính này được thiết kế dưới dạng thang điểm từ 1 đến 10, đại diện cho mức độ căng thẳng của người dùng. Việc dự đoán *stress_level* dựa trên các thuộc tính còn lại giúp mô hình hóa mối quan hệ giữa hành vi công nghệ và tình trạng stress, đồng thời hỗ trợ quá trình phân cụm người dùng theo các nhóm đặc điểm tương đồng.

Tổng thể, bộ dữ liệu với hệ thống thuộc tính đa dạng và có giá trị mô tả cao đã tạo điều kiện thuận lợi cho việc triển khai các kỹ thuật khai phá dữ liệu trong đồ án, bao gồm phân tích thống kê mô tả, dự đoán mức độ căng thẳng bằng cây quyết định và phân nhóm người dùng bằng thuật toán phân cụm.

1.6 Giới thiệu các công cụ sử dụng trong đồ án

Trong quá trình triển khai đồ án *"Ứng dụng cây quyết định và thuật toán phân cụm trong đánh giá tác động của công nghệ đối với tâm lý con người"*, việc lựa chọn công cụ lập trình, môi trường phát triển và nền tảng thực thi đóng vai trò quan trọng trong việc đảm bảo tính chính xác, hiệu quả và khả năng tái lập của các thí nghiệm. Các công cụ được sử dụng không chỉ hỗ trợ thao tác trên dữ liệu mà còn giúp trực quan hóa quá trình phân tích, tối ưu mô hình và trình bày kết quả một cách khoa học. Ba công cụ

chính được sử dụng trong đồ án bao gồm: Python, Visual Studio Code và Google Colab. Nội dung dưới đây trình bày chi tiết vai trò và đặc điểm của từng công cụ.

1.6.1 Giới thiệu về Python

Python là ngôn ngữ lập trình trung tâm được sử dụng xuyên suốt toàn bộ đồ án, nhờ những đặc tính nổi bật về sự linh hoạt, cú pháp rõ ràng và khả năng hỗ trợ mạnh mẽ cho lĩnh vực khoa học dữ liệu. Python hiện là một trong những ngôn ngữ được lựa chọn hàng đầu trong các lĩnh vực như phân tích dữ liệu, trí tuệ nhân tạo, học máy và khai phá dữ liệu nhờ hệ sinh thái thư viện hết sức phong phú.

Các thư viện cốt lõi như **NumPy** và **Pandas** giúp xử lý, biến đổi và phân tích dữ liệu với tốc độ nhanh và thao tác dễ dàng; **Matplotlib** và **Seaborn** hỗ trợ trực quan hóa dữ liệu thông qua biểu đồ, giúp người nghiên cứu nhận diện các mẫu hành vi và mối quan hệ tiềm ẩn giữa các biến. Trong khi đó, **Scikit-Learn** cung cấp nhiều thuật toán học máy, bao gồm cây quyết định (Decision Tree), phân cụm K-Means, lựa chọn đặc trưng, chuẩn hóa dữ liệu và các phương pháp đánh giá mô hình.

Bên cạnh hệ thống thư viện đa dạng, Python còn được đánh giá cao nhờ cộng đồng người dùng rộng lớn và tài liệu hỗ trợ phong phú, tạo điều kiện thuận lợi để giải quyết các vấn đề trong quá trình thực hiện đồ án. Ngôn ngữ này cũng phù hợp cho các nghiên cứu mang tính thực nghiệm nhờ khả năng thử nghiệm nhanh các thuật toán và điều chỉnh tham số mô hình một cách linh hoạt.

Nhờ những ưu điểm trên, Python trở thành công cụ cốt lõi hỗ trợ mọi giai đoạn của đồ án, từ tiền xử lý dữ liệu, xây dựng mô hình đến đánh giá và trực quan hóa kết quả.

1.6.2 Giới thiệu về Visual Studio Code

Visual Studio Code (VS Code) được sử dụng làm môi trường phát triển chính để tổ chức mã nguồn và triển khai các thuật toán trong đồ án. VS Code là một trình soạn thảo mã nguồn mạnh mẽ, nhẹ và linh hoạt, được phát triển bởi Microsoft. Công cụ này hỗ trợ nhiều ngôn ngữ lập trình khác nhau, trong đó có Python – ngôn ngữ được sử dụng trong đồ án.

Nhờ kho tiện ích mở rộng phong phú, VS Code cung cấp nhiều công cụ quan trọng phục vụ cho quá trình lập trình như hệ thống gợi ý mã thông minh, kiểm tra lỗi cú pháp theo thời gian thực, chạy mã nhanh thông qua Code Runner, và tích hợp sâu với

Jupyter Notebook để thực thi các đoạn mã Python theo từng ô lệnh. Bên cạnh đó, VS Code còn cho phép cài đặt và quản lý môi trường ảo Python, giúp tách biệt phiên bản thư viện cho từng dự án, đảm bảo tính ổn định và đồng bộ trong quá trình phát triển.

Công cụ này còn hỗ trợ hệ thống quản lý phiên bản Git, giúp theo dõi các thay đổi của mã nguồn, lưu trữ lịch sử chỉnh sửa và phối hợp làm việc nhóm một cách hiệu quả, dù đồ án thực hiện cá nhân hay theo nhóm. Tính năng giao diện tối giản và dễ tùy chỉnh cũng giúp VS Code thích hợp cho các dự án dài hạn.

Nhờ những ưu điểm trên, VS Code được lựa chọn làm nền tảng lập trình chính để xây dựng, thử nghiệm và tối ưu mã nguồn trong đồ án.

1.6.3 Giới thiệu về Google Colab

Google Colab (Colaboratory) là nền tảng tính toán dựa trên đám mây do Google phát triển, cho phép chạy mã Python trực tiếp trên trình duyệt mà không yêu cầu cài đặt môi trường lập trình tại máy tính cá nhân. Công cụ này đặc biệt hữu ích trong các bài toán khoa học dữ liệu nhờ khả năng cung cấp tài nguyên tính toán miễn phí, bao gồm CPU, GPU và TPU.

Colab hỗ trợ định dạng notebook (.ipynb), một dạng tài liệu tương tác kết hợp giữa mã nguồn, văn bản mô tả và kết quả trực quan hóa. Điều này giúp cho việc ghi chú, trình bày thí nghiệm và theo dõi các bước xử lý dữ liệu trở nên rõ ràng hơn. Hệ thống thư viện Python phổ biến được cài đặt sẵn, giúp người nghiên cứu tiết kiệm thời gian thiết lập môi trường.

Một trong những ưu điểm quan trọng của Colab là khả năng liên kết trực tiếp với Google Drive, cho phép lưu trữ dữ liệu và mô hình trong môi trường đám mây, dễ dàng chia sẻ notebook với người khác và đảm bảo tính tái lập của quá trình thực nghiệm. Ngoài ra, người dùng có thể nhanh chóng mở rộng dung lượng lưu trữ hoặc tải lên các dataset lớn mà không bị giới hạn dung lượng như trên máy tính cá nhân.

Trong đồ án này, Google Colab được sử dụng như môi trường hỗ trợ quá trình kiểm thử mô hình, đánh giá thuật toán và thực hiện các tác vụ tính toán cần đến tài nguyên lớn. Đây cũng là công cụ thuận tiện để trực quan hóa kết quả phân tích và ghi lại tiến trình thực nghiệm theo từng bước.

CHƯƠNG 2: THUẬT TOÁN KHAI THÁC DỮ LIỆU SỮ DỤNG

2.1 Sử dụng thuật toán Decision Tree

Trong đồ án này, thuật toán Cây Quyết Định (Decision Tree) được sử dụng như một phương pháp phân lớp nhằm dự đoán mức độ căng thẳng (*stress_level*) của người dùng dựa trên các thuộc tính liên quan đến hành vi sử dụng công nghệ và các yếu tố tâm lý – lối sống. Cây Quyết Định là một trong những thuật toán trực quan và dễ giải thích nhất trong học máy, đặc biệt phù hợp cho các bài toán yêu cầu khả năng giải thích mô hình (model interpretability).

Việc áp dụng Decision Tree trong đồ án giúp xây dựng một mô hình có thể mô tả rõ ràng những yếu tố hoặc điều kiện đặc trưng dẫn đến các mức độ căng thẳng khác nhau. Thông qua cấu trúc phân nhánh của cây, người đọc có thể nhận biết những biến đầu vào quan trọng, chẳng hạn như thời gian sử dụng mạng xã hội, thời gian làm việc trên thiết bị, thời lượng ngủ hoặc các chỉ số lo âu – trầm cảm. Tính minh bạch trong quá trình ra quyết định của thuật toán giúp người nghiên cứu dễ dàng phân tích tác động của từng thuộc tính, từ đó hỗ trợ cho việc đưa ra nhận định chính xác về mối quan hệ giữa công nghệ và sức khỏe tinh thần.

Bên cạnh đó, Decision Tree còn là nền tảng cho nhiều phương pháp học máy tiên tiến như Random Forest và Gradient Boosting, do đó việc hiểu rõ nguyên lý hoạt động của mô hình này giúp củng cố kiến thức nền tảng cho các thuật toán phức tạp hơn. Trong phạm vi đồ án, Decision Tree được sử dụng như một thuật toán độc lập để thực hiện nhiệm vụ phân lớp mức độ stress, đồng thời phục vụ mục tiêu giải thích mô hình và đánh giá mức độ quan trọng của các thuộc tính dữ liệu.

2.1.1 Tổng quan về thuật toán phân lớp dựa trên Decision Tree

Thuật toán Decision Tree là một phương pháp học có giám sát, hoạt động dựa trên việc xây dựng một cấu trúc dạng cây nhằm phân chia dữ liệu thành các nhóm nhỏ hơn dựa trên các thuộc tính đặc trưng. Mục tiêu của quá trình phân chia là tạo ra các nút lá biểu diễn các lớp đầu ra thuần nhất nhất có thể. Trong bài toán phân lớp, Decision Tree xác định chuỗi các câu hỏi dạng điều kiện dựa trên giá trị của các thuộc tính để quyết định mẫu dữ liệu thuộc về lớp nào.

Cấu trúc cơ bản của một cây quyết định bao gồm ba thành phần: **nút gốc**, **nút trung gian** và **nút lá**. Nút gốc là điểm bắt đầu của cây, nơi đặt điều kiện đầu tiên chia dữ liệu thành hai hoặc nhiều nhánh. Nút trung gian tiếp tục thực hiện các phép chia tương tự dựa trên các thuộc tính còn lại. Cuối cùng, các nút lá chứa nhãn phân lớp và là nơi mô hình đưa ra dự đoán.

Quy trình xây dựng Decision Tree được thực hiện thông qua việc lựa chọn thuộc tính tối ưu để phân chia dữ liệu tại mỗi bước. Việc lựa chọn này dựa trên các thước đo độ thuần nhất (impurity measures), phổ biến nhất là **Entropy**, **Information Gain**, **Gini Index** và đôi khi là **Chi-square**. Mục tiêu của thuật toán là tìm ra thuộc tính giúp phân tách dữ liệu thành các nhóm thuần nhất nhất, tức là các nhóm có biến mục tiêu gần như giống nhau.

Decision Tree có nhiều ưu điểm nổi bật. Trước hết, mô hình rất dễ hiểu và dễ trình bày, phù hợp cho các bài toán cần tính giải thích cao. Người dùng có thể quan sát trực tiếp cây phân cấp để hiểu được cách mô hình đưa ra quyết định. Ngoài ra, Decision Tree có khả năng xử lý cả dữ liệu số và dữ liệu phân loại mà không cần chuẩn hóa thang đo, đồng thời xử lý tốt các mối quan hệ phi tuyến tính giữa các biến.

Tuy nhiên, thuật toán cũng tồn tại một số hạn chế, chẳng hạn như dễ bị overfitting nếu cây phát triển quá sâu, nhạy cảm với nhiễu và dễ thay đổi cấu trúc khi dữ liệu đầu vào biến động nhẹ. Để khắc phục điều này, các kỹ thuật cắt tỉa cây (pruning) và giới hạn độ sâu thường được áp dụng để cải thiện khả năng tổng quát hóa của mô hình.

Trong bối cảnh bài toán của đồ án, Decision Tree là lựa chọn phù hợp bởi nó vừa phục vụ tốt mục tiêu phân lớp mức độ stress, vừa mang lại khả năng giải thích rõ ràng về ảnh hưởng của từng yếu tố công nghệ và tâm lý đối với tình trạng căng thẳng của người dùng.

2.1.2 Lý do chọn thuật toán Decision Tree

Lựa chọn thuật toán Cây Quyết Định (Decision Tree) trong đồ án xuất phát từ nhiều đặc điểm nổi bật của phương pháp này, đặc biệt phù hợp với bài toán dự đoán mức độ căng thẳng dựa trên dữ liệu hành vi sử dụng công nghệ và các yếu tố tâm lý – lối sống. Decision Tree không chỉ là một thuật toán mạnh mẽ trong phân lớp mà còn mang lại khả năng diễn giải cao, điều này giúp người nghiên cứu không chỉ tập trung

vào kết quả dự đoán mà còn hiểu rõ cơ chế tác động của từng thuộc tính dữ liệu đến biến mục tiêu.

Một trong những lý do quan trọng nhất để lựa chọn Decision Tree là tính minh bạch trong quá trình ra quyết định. Không giống như nhiều thuật toán học máy khác hoạt động như "hộp đen", Decision Tree thể hiện rõ ràng các điều kiện phân chia dữ liệu thông qua cấu trúc dạng cây. Điều này giúp người đọc dễ dàng quan sát được các yếu tố như thời gian sử dụng mạng xã hội, thời lượng ngủ, mức độ lo âu hay thời gian làm việc trên thiết bị tác động như thế nào đến mức độ stress. Trong bối cảnh bài toán liên quan đến sức khỏe tinh thần, khả năng giải thích này đặc biệt cần thiết để cung cấp các luận giải hợp lý, làm cơ sở cho việc đưa ra các khuyến nghị hay đánh giá khoa học.

Bên cạnh tính dễ giải thích, Decision Tree còn có khả năng xử lý dữ liệu hỗn hợp gồm cả biến liên tục và biến phân loại mà không yêu cầu chuẩn hóa thang đo. Điều này rất phù hợp với dataset của đồ án, nơi tồn tại nhiều loại dữ liệu khác nhau như số giờ sử dụng thiết bị (dạng số thực), mức độ lo âu và trầm cảm (dạng chỉ số), giới tính và thói quen ăn uống (dạng phân loại). Việc không phải thực hiện quá nhiều bước chuẩn hóa giúp tiết kiệm thời gian tiền xử lý và giảm nguy cơ làm mất thông tin quan trọng trong dữ liệu.

Một ưu điểm khác của Decision Tree là khả năng mô hình hóa các mối quan hệ phi tuyến tính. Trong thực tế, mức độ stress của con người thường không biến đổi tuyến tính theo từng yếu tố riêng lẻ mà chịu ảnh hưởng bởi nhiều điều kiện kết hợp. Decision Tree có thể biểu diễn những mối quan hệ phức tạp này thông qua các nhánh phân chia, nhờ đó mô hình có thể đạt độ chính xác cao hơn trong những trường hợp dữ liệu mang tính tương tác mạnh giữa các biến.

Ngoài ra, tốc độ huấn luyện của Decision Tree tương đối nhanh so với nhiều thuật toán phân lớp khác. Điều này tạo thuận lợi cho việc thử nghiệm, điều chỉnh tham số và đánh giá mô hình trong nhiều vòng lặp, đặc biệt khi số lượng quan sát lớn như dữ liệu 5000 mẫu trong đồ án. Mô hình cũng cho phép đánh giá mức độ quan trọng của từng thuộc tính (feature importance), giúp người nghiên cứu xác định được những yếu tố có ảnh hưởng mạnh nhất đến stress, qua đó phục vụ mục tiêu phân tích tác động của công nghệ đến tâm lý con người.

Mặc dù tồn tại những hạn chế như nguy cơ overfitting khi cây quá phức tạp, thuật toán vẫn phù hợp với mục tiêu đồ án khi kết hợp cùng các kỹ thuật như giới hạn độ sâu cây hoặc cắt tỉa (pruning). Việc điều chỉnh này giúp tăng khả năng tổng quát hóa của mô hình mà vẫn giữ nguyên ưu điểm về tính trực quan.

2.1.3 Tập huấn luyện (Train Set) của thuật toán Decision Tree

Trong quá trình xây dựng mô hình phân lớp bằng thuật toán Decision Tree, dữ liệu được chia thành hai phần bao gồm tập huấn luyện (train set) và tập kiểm thử (test set). Việc phân chia này nhằm đảm bảo mô hình được đào tạo trên một tập dữ liệu đủ lớn và đa dạng, từ đó học được các quy luật cần thiết để dự đoán mức độ căng thẳng (*stress_level*) dựa trên các thuộc tính đầu vào. Đồng thời, cấu trúc tách dữ liệu cũng giúp đánh giá khách quan khả năng tổng quát hóa của mô hình trên dữ liệu chưa từng gặp trước đó.

Trong chương trình của đồ án, quá trình tách dữ liệu được thực hiện bằng hàm **train_test_split** thuộc thư viện *Scikit-Learn*. Tập dữ liệu ban đầu sau khi hoàn thiện tiền xử lý được chia theo tỉ lệ 80% dùng cho huấn luyện và 20% dùng cho kiểm thử. Phần train set bao gồm các mẫu dữ liệu được mô hình sử dụng để xây dựng cấu trúc cây bằng cách lựa chọn thuộc tính phân chia tối ưu tại mỗi nút. Thuật toán tiến hành phân tích mối quan hệ giữa các thuộc tính như thời gian sử dụng mạng xã hội, thời lượng ngủ, mức độ lo âu, thời gian làm việc trên thiết bị và nhiều yếu tố khác để học được các quy tắc phân lớp.

Ngoài việc sử dụng để xây dựng cấu trúc mô hình, tập huấn luyện còn đóng vai trò quan trọng trong việc xác định các giá trị siêu tham số (hyperparameters) như độ sâu tối đa của cây, số lượng mẫu tối thiểu tại mỗi nút phân chia hay số lượng mẫu tối thiểu để tạo thành một nút lá. Việc điều chỉnh hợp lý các tham số này dựa trên train set giúp hạn chế nguy cơ overfitting, đồng thời nâng cao độ chính xác của mô hình trên dữ liệu thực tế..

age	gender	daily_screen_time_hours	sleep_duration_hours	social_media_hours	work_related_hours	gaming_hours	phone_usage_hours
0.423728814	0	0.533333333	0.555555556	0.928571429	0.736842105	0.315789474	0.520833333
0.542372881	0	0.188888889	0.583333333	0.071428571	0.184210526	0.473684211	0.375
0.711864407	0.5	0.377777778	0.527777778	1	0.894736842	0.368421053	0.416666667
0.745762712	0	0.422222222	0.666666667	0.666666667	0.868421053	0.447368421	0.479166667
0.152542373	0	0.744444444	0.222222222	0.857142857	0.815789474	0.289473684	0.3125
0.372881356	0.5	0.477777778	0.611111111	0.880952381	0.815789474	0.157894737	0.5
0.610169492	0.5	0.522222222	0.527777778	1	0.894736842	0.342105263	0.770833333
0.576271186	0.5	0.011111111	0.666666667	0.880952381	0.815789474	0.289473684	0.291666667
0.237288136	0.5	0.588888889	0.666666667	0.833333333	0.921052632	0.210526316	0.458333333
0.813559322	1	0.633333333	0.361111111	0.952380952	0.815789474	0.421052632	0.583333333
0.847457627	0	0.711111111	0.277777778	1	0.842105263	0.210526316	0.208333333
0.610169492	1	0.077777778	0.555555556	0.428571429	0.552631579	0.526315789	0.1875
0.728813559	0	0.322222222	0.305555556	0.595238095	0.736842105	0.315789474	0.3125
0.525423729	0.5	0.511111111	0.5	0.476190476	0.736842105	0.368421053	0.520833333
0.610169492	0.5	0.533333333	0.611111111	1	0.921052632	0.394736842	0.395833333
0.050847458	0	0.544444444	0.472222222	1	0.947368421	0.263157895	0.3125
0.355932203	0.5	0.822222222	0.555555556	1	0.842105263	0.552631579	0.395833333
0.169491525	0.5	0.144444444	0.527777778	0.619047619	0.736842105	0.473684211	0.041666667
0.271186441	0	0.5	0.277777778	0.785714286	0.894736842	0.368421053	0.583333333
0.423728814	1	0.5	0.416666667	0.880952381	0.894736842	0.236842105	0.3125
0.338983051	0.5	0.133333333	0.833333333	0.142857143	0.210526316	0.710526316	0.541666667
0.762711864	0	0.255555556	0.527777778	0.976190476	0.947368421	0.421052632	0.291666667
0.559322034	0.5	0.266666667	0.5	0.595238095	0.605263158	0.447368421	0.520833333
0.644067797	0	0.277777778	0.555555556	0.476190476	0.605263158	0.421052632	0.375
0.457627119	1	0.6	0.277777778	1	0.815789474	0.210526316	0.583333333
0.898305085	1	0.444444444	0.638888889	0.80952381	0.815789474	0.342105263	0.25
0.898305085	0	0.455555556	0.472222222	0.952380952	0.842105263	0.210526316	0.270833333
0.06779661	0	0.322222222	0.888888889	0.761904762	0.815789474	0.447368421	0
0.525423729	0.5	0.155555556	0.777777778	0.523809524	0.710526316	0.736842105	0.0625
0.254237288	0.5	0.644444444	0.638888889	1	0.894736842	0.184210526	0.416666667
0.491525424	0.5	0.422222222	0.722222222	1	0.842105263	0.131578947	0.625
0.237288136	0.5	0.388888889	0.361111111	0.80952381	0.868421053	0.157894737	0.520833333
0.644067797	0.5	0.177777778	0.583333333	0.119047619	0.184210526	0.631578947	0.166666667
0.050847458	0	0.477777778	0.694444444	1	0.815789474	0.368421053	0.0625
1	0.5	0.711111111	0.444444444	0.785714286	0.710526316	0.184210526	0.479166667
0.186440678	0	0.7	0.694444444	1	0.868421053	0.263157895	0.479166667
0.440677966	0	0.477777778	0.583333333	1	0.894736842	0.105263158	0.729166667
0.423728814	0	0.544444444	0.416666667	0.5	0.631578947	0.710526316	0.208333333

Hình 2. 1 Tập data train decision-tree 1

laptop_usage_hours	sleep_quality	health_score	sleep_health_index	emotional_balance	overall_wellness	digital_stress_score	work_life_balance
0.78	0.75	0.579710145	0.617021277	0.3	0.454545455	0.658862876	0.263157895
0.18	1	0.797101449	0.840425532	0.988888889	0.922459893	0.205685619	0.815789474
0.12	0.75	0.260869565	0.606382979	0.077777778	0.227272727	0.585284281	0.105263158
0.26	1	0.536231884	0.872340426	0.366666667	0.534759358	0.519230769	0.131578947
0.32	0.75	0.362318841	0.489361702	0.277777778	0.331550802	0.640468227	0.184210526
0.36	0.75	0.376811594	0.638297872	0.111111111	0.294117647	0.615384615	0.184210526
0.26	0.75	0.289855072	0.606382979	0.033333333	0.21657754	0.767558528	0.105263158
0.02	1	0.434782609	0.872340426	0.366666667	0.497326203	0.364548495	0.184210526
0	0.75	0.623188406	0.659574468	0.311111111	0.486631016	0.627090301	0.078947368
0.44	0.5	0.391304348	0.329787234	0	0.168449198	0.726588629	0.184210526
0.62	0.75	0.289855072	0.510638298	0.022222222	0.187165775	0.635451505	0.157894737
0	1	0.797101449	0.829787234	0.722222222	0.79144385	0.2090301	0.447368421
0.34	1	0.695652174	0.734042553	0.566666667	0.655080214	0.398829431	0.263157895
0.52	0.75	0.391304348	0.595744681	0.511111111	0.481283422	0.507525084	0.263157895
0.28	0.5	0.391304348	0.425531915	0.044444444	0.213903743	0.636287625	0.078947368
0.24	0.5	0.202898551	0.372340426	0	0.109625668	0.610367893	0.052631579
0.6	0.5	0.217391304	0.404255319	0	0.122994652	0.744983278	0.157894737
0.08	1	0.565217391	0.819148936	0.577777778	0.63368984	0.241638796	0.263157895
0.08	0.75	0.304347826	0.510638298	0.333333333	0.342245989	0.623745819	0.105263158
0.4	0.75	0.376811594	0.563829787	0.222222222	0.328877005	0.556020067	0.105263158
0.14	1	0.739130435	0.936170213	0.811111111	0.839572193	0.267558528	0.789473684
0.24	0.75	0.130434783	0.606382979	0.055555556	0.168449198	0.486622074	0.052631579
0.1	0.75	0.695652174	0.595744681	0.522222222	0.598930481	0.453177258	0.394736842
0	1	0.782608696	0.829787234	0.577777778	0.71657754	0.367056856	0.394736842
0.5	0.75	0.188405797	0.510638298	0	0.139037433	0.72909699	0.184210526
0.3	0.75	0.710144928	0.64893617	0.333333333	0.526737968	0.489966555	0.184210526
0.12	0.75	0.47826087	0.585106383	0.255555556	0.387700535	0.546822742	0.157894737
0.32	1	0.420289855	0.957446809	0.233333333	0.449197861	0.338628763	0.184210526
0.34	0.75	0.652173913	0.70212766	0.533333333	0.614973262	0.223244147	0.289473684
0.34	0.5	0.304347826	0.436170213	0	0.163101604	0.685618729	0.105263158
0.2	0.75	0.333333333	0.680851064	0	0.235294118	0.677257525	0.157894737
0	0.75	0.47826087	0.542553191	0.4	0.446524064	0.566889632	0.131578947
0.2	1	0.753623188	0.840425532	0.777777778	0.804812834	0.141304348	0.815789474
0.28	0.75	0.057971014	0.670212766	0	0.131016043	0.494983278	0.184210526
0.5	0.5	0.652173913	0.361702128	0.422222222	0.475935829	0.665551839	0.289473684
0.32	0.5	0.333333333	0.457446809	0.1	0.227272727	0.72909699	0.131578947
0.18	0.75	0.304347826	0.627659574	0.033333333	0.227272727	0.735785953	0.105263158
0.56	0.75	0.594202899	0.563829787	0.588888889	0.585561497	0.414715719	0.368421053

Hình 2. 2 Tập data train decision-tree 2

2.1.4 Tập kiểm thử (Test Set) của thuật toán Decision Tree

Tập kiểm thử (test set) được tách ra như một phần độc lập với tập huấn luyện nhằm đánh giá hiệu suất tổng quát của mô hình Decision Tree sau khi quá trình đào tạo hoàn tất. Trong chương trình của đồ án, test set chiếm 20% tổng số mẫu dữ liệu và không được mô hình sử dụng trong bất kỳ giai đoạn huấn luyện nào. Điều này đảm bảo rằng việc đánh giá mô hình phản ánh đúng khả năng dự đoán của thuật toán trên dữ liệu mới, giúp đo lường mức độ tin cậy của mô hình trong thực tế.

Sau khi mô hình Decision Tree được huấn luyện bằng train set, test set được đưa vào để kiểm tra khả năng phân lớp mức độ stress. Dựa trên các thuộc tính của từng quan sát trong test set, mô hình dự đoán giá trị stress_level và kết quả dự đoán được so sánh với giá trị thực tế để tính toán các chỉ số đánh giá như độ chính xác (accuracy), độ nhạy

(recall), độ đặc hiệu (precision) hoặc ma trận nhầm lẫn (confusion matrix). Các chỉ số này phản ánh mức độ phù hợp của mô hình đối với bài toán và chỉ ra những điểm mạnh, điểm hạn chế của thuật toán khi áp dụng vào dữ liệu chưa biết trước.

Việc đánh giá mô hình bằng test set giúp xác định mức độ tin cậy của Decision Tree và kiểm tra xem mô hình có bị overfitting hoặc underfitting hay không. Nếu độ chính xác trên tập huấn luyện cao nhưng trên test set giảm mạnh, điều này cho thấy mô hình chưa có khả năng tổng quát hóa tốt và cần điều chỉnh cấu trúc cây hoặc siêu tham số. Ngược lại, nếu mô hình đạt hiệu suất ổn định giữa hai tập dữ liệu, điều đó chứng tỏ Decision Tree đã học được những quy luật phù hợp và có thể vận dụng hiệu quả vào các dự đoán thực tế

age	gender	daily_screen_time_hours	sleep_duration_hours	social_media_hours	work_related_hours	gaming_hours	phone_usage_hours
0.576271186	0	0.577777778	0.416666667	1	0.921052632	0.473684211	0.395833333
0.949152542	0.5	0.344444444	0.972222222	0.547619048	0.657894737	0.605263158	0.229166667
0.898305085	0	0.644444444	0.222222222	0.80952381	0.815789474	0.184210526	0.645833333
0.525423729	0.5	0.711111111	0.611111111	1	0.815789474	0.263157895	0.645833333
0.542372881	0	0.433333333	0.638888889	0.404761905	0.421052632	0.605263158	0.3125
0.389830508	0	0.677777778	0.583333333	1	0.842105263	0.105263158	0.583333333
0.406779661	0.5	0.555555556	0.555555556	1	0.815789474	0.236842105	0.645833333
0.474576271	0.5	0.522222222	0.583333333	1	0.921052632	0.315789474	0.4375
0.847457627	0	0.588888889	0.416666667	0.904761905	0.815789474	0.289473684	0.375
0.559322034	0.5	0.411111111	0.277777778	0.642857143	0.684210526	0.447368421	0.375
0.593220339	0	0.577777778	0.666666667	1	0.921052632	0.473684211	0.291666667
0.779661017	0.5	0.333333333	0.694444444	0.357142857	0.526315789	0.578947368	0.1875
0.949152542	0	0.2	0.611111111	0.142857143	0.131578947	0.763157895	0.104166667
0.576271186	0.5	0.222222222	0.694444444	0.095238095	0.184210526	0.605263158	0.208333333
0.881355932	0.5	0.544444444	0.638888889	0.904761905	0.842105263	0.289473684	0.416666667
0.389830508	0.5	0.277777778	0.5	0.142857143	0.131578947	0.815789474	0.270833333
0.93220339	0	0.555555556	0.611111111	0.738095238	0.789473684	0.394736842	0.416666667
0.288135593	0	0.377777778	0.277777778	0.904761905	0.842105263	0.263157895	0.229166667
0.491525424	0.5	0.733333333	0.472222222	0.80952381	0.763157895	0.263157895	0.395833333
0.271186441	0	0.422222222	0.694444444	0.833333333	0.815789474	0.315789474	0.4375
0.050847458	0	0.944444444	0.5	1	0.921052632	0.184210526	0.708333333
0.86440678	0	0.344444444	0.611111111	0.547619048	0.763157895	0.473684211	0.416666667
0.406779661	0.5	0.322222222	0.555555556	0.761904762	0.815789474	0.421052632	0.354166667
0.508474576	0.5	0.577777778	0.25	1	0.789473684	0.421052632	0.666666667
0.610169492	0	0.233333333	0.694444444	0.142857143	0.052631579	0.815789474	0.3125
0.220338983	1	0.266666667	0.638888889	0.357142857	0.368421053	0.684210526	0.416666667
0.101694915	0.5	0.633333333	0.555555556	1	0.894736842	0.157894737	0.291666667
0.644067797	0.5	0.288888889	0.555555556	0.714285714	0.789473684	0.368421053	0.4375
0.457627119	0	0.733333333	0.75	1	0.815789474	0.078947368	0.479166667
0.610169492	0.5	0.544444444	0.583333333	0.80952381	0.789473684	0.368421053	0.604166667
0.440677966	0.5	0	0.694444444	0.857142857	0.868421053	0.447368421	0.125
0.983050847	0	0.588888889	0.555555556	0.333333333	0.473684211	0.605263158	0.416666667
0.186440678	0.5	0.566666667	0.416666667	1	0.789473684	0.289473684	0.375
0.254237288	0	0.333333333	0.472222222	0.714285714	0.815789474	0.289473684	0.520833333
0.305084746	0.5	0.333333333	0.527777778	0.666666667	0.736842105	0.552631579	0.4375
0.355932203	0	0.3	0.5	0.666666667	0.736842105	0.394736842	0.666666667
0.389830508	0	0.4	0.527777778	1	0.842105263	0.263157895	0.145833333
0.644067797	0	0.288888889	0.305555556	0.761904762	0.842105263	0.394736842	0.125

Hình 2. 3 Tập data test decision-tree 1

laptop_usage_hours	sleep_quality	health_score	sleep_health_index	emotional_balance	overall_wellness	digital_stress_score	work_life_balance
0.46	0.5	0.347826087	0.35106383	0	0.157754011	0.653010033	0.078947368
0.24	1	0.550724638	0.989361702	0.677777778	0.719251337	0.362040134	0.342105263
0.36	0.75	0.565217391	0.489361702	0.3	0.417112299	0.70819398	0.184210526
0.24	0.75	0.449275362	0.638297872	0.177777778	0.352941176	0.793478261	0.184210526
0.14	0.75	0.724637681	0.64893617	0.722222222	0.719251337	0.380434783	0.578947368
0.54	0.5	0.173913043	0.414893617	0	0.109625668	0.758361204	0.157894737
0.58	0.75	0.391304348	0.617021277	0.022222222	0.251336898	0.734949833	0.184210526
0.66	0.75	0.304347826	0.627659574	0	0.211229947	0.647157191	0.078947368
0.38	0.75	0.47826087	0.563829787	0.288888889	0.398395722	0.619565217	0.184210526
0.1	1	0.666666667	0.723404255	0.6	0.657754011	0.469899666	0.315789474
0.24	1	0.434782609	0.872340426	0.111111111	0.374331551	0.615384615	0.078947368
0.5	0.75	0.608695652	0.670212766	0.8	0.719251337	0.282608696	0.473684211
0.34	1	0.710144928	0.85106383	0.844444444	0.823529412	0.134615385	0.868421053
0.04	1	0.797101449	0.882978723	0.9	0.890374332	0.165551839	0.815789474
0.3	0.75	0.362318841	0.64893617	0.288888889	0.377005348	0.617892977	0.157894737
0.36	0.75	0.666666667	0.595744681	1	0.818181818	0.224080268	0.868421053
0.46	1	0.724637681	0.85106383	0.511111111	0.668449198	0.569397993	0.210526316
0.68	0.75	0.347826087	0.510638298	0.177777778	0.28342246	0.487458194	0.157894737
0.44	0.75	0.420289855	0.585106383	0.311111111	0.393048128	0.651337793	0.236842105
0.36	0.75	0.492753623	0.670212766	0.155555556	0.36631016	0.556856187	0.184210526
0.46	0.5	0.31884058	0.382978723	0	0.155080214	0.903846154	0.078947368
0.62	0.75	0.52173913	0.638297872	0.5	0.534759358	0.429765886	0.236842105
0.38	1	0.608695652	0.829787234	0.455555556	0.593582888	0.466555184	0.184210526
0.22	0.5	0.260869565	0.287234043	0	0.109625668	0.75083612	0.210526316
0.3	1	0.898550725	0.882978723	1	0.975935829	0.222408027	0.947368421
0.22	0.75	0.797101449	0.64893617	0.855555556	0.810160428	0.340301003	0.631578947
0.12	0.75	0.188405797	0.617021277	0	0.165775401	0.636287625	0.105263158
0.14	0.75	0.463768116	0.617021277	0.422222222	0.470588235	0.469063545	0.210526316
0.64	0.75	0.31884058	0.691489362	0.066666667	0.264705882	0.741638796	0.184210526
0.46	0.75	0.463768116	0.627659574	0.144444444	0.339572193	0.655518395	0.210526316
0.02	0.75	0.507246377	0.670212766	0.233333333	0.409090909	0.29264214	0.131578947
0.6	0.75	0.695652174	0.617021277	0.788888889	0.732620321	0.454013378	0.526315789
0.42	0.75	0.405797101	0.563829787	0.077777778	0.270053476	0.641304348	0.210526316
0.28	1	0.565217391	0.79787234	0.444444444	0.564171123	0.515886288	0.184210526
0.26	0.75	0.623188406	0.606382979	0.455555556	0.542780749	0.470735786	0.263157895
0.16	0.75	0.608695652	0.595744681	0.533333333	0.572192513	0.5409699	0.263157895
0.52	0.75	0.173913043	0.606382979	0.255555556	0.280748663	0.495819398	0.157894737
0.64	0.75	0.362318841	0.521276596	0.444444444	0.419786096	0.371237458	0.157894737

Hình 2. 4 Tập data test decision-tree 2

2.2 Sử dụng thuật toán Random Forest

Trong đồ án này, thuật toán Random Forest được sử dụng như một phương pháp phân lớp bổ sung nhằm nâng cao độ chính xác trong việc dự đoán mức độ căng thẳng (*stress_level*) của người dùng. Random Forest là một thuật toán học máy thuộc nhóm mô hình tổ hợp (ensemble learning), hoạt động dựa trên ý tưởng kết hợp nhiều cây quyết định yếu thành một mô hình mạnh. Cách tiếp cận này đã chứng minh hiệu quả vượt trội so với việc sử dụng một cây quyết định đơn lẻ, đặc biệt trong những bài toán có dữ liệu nhiễu, biến đầu vào đa dạng hoặc quan hệ phi tuyến phức tạp.

Việc đưa Random Forest vào đồ án không chỉ nhằm mục tiêu so sánh hiệu suất giữa các thuật toán phân lớp mà còn nhằm nâng cao tính tin cậy của mô hình dự đoán. Thuật toán này có khả năng giảm thiểu hiện tượng overfitting vốn có trong Decision Tree, nhờ vào việc tạo ra nhiều cây với sự ngẫu nhiên trong quá trình lựa chọn dữ liệu

và thuộc tính. Sự kết hợp của nhiều mô hình con giúp Random Forest đạt được mức ổn định cao hơn, đồng thời cung cấp thông tin quan trọng về mức độ đóng góp của từng thuộc tính, hỗ trợ đắc lực cho việc phân tích mối quan hệ giữa hành vi công nghệ và tình trạng stress của người dùng.

Việc sử dụng Random Forest trong đồ án góp phần làm rõ hơn những yếu tố quan trọng ảnh hưởng đến mức độ căng thẳng và tăng cường sự chính xác trong dự đoán, đồng thời nâng cao tính toàn diện của quá trình phân tích dữ liệu.

2.2.1 Tổng quan về thuật toán phân lớp dựa trên Random Forest

Random Forest là một thuật toán phân lớp và hồi quy mạnh mẽ, được phát triển dựa trên phương pháp Bagging (Bootstrap Aggregating), kết hợp với cơ chế lựa chọn thuộc tính ngẫu nhiên nhằm tạo ra một tập hợp lớn các cây quyết định (decision trees). Mỗi cây trong rừng được xây dựng từ một mẫu dữ liệu bootstrap – tức là một tập con dữ liệu được lấy mẫu ngẫu nhiên có hoàn lại từ bộ dữ liệu gốc. Bên cạnh đó, tại mỗi nút phân chia của cây, thuật toán chỉ xem xét một tập con nhỏ các thuộc tính được chọn ngẫu nhiên thay vì toàn bộ các thuộc tính. Hai yếu tố ngẫu nhiên này giúp tạo ra sự đa dạng giữa các cây trong mô hình, nhờ đó tăng khả năng tổng quát hóa và giảm độ nhạy với nhiễu.

Khi tiến hành dự đoán, Random Forest tổng hợp kết quả dự đoán từ tất cả các cây trong mô hình thông qua phương pháp bỏ phiếu đa số (*majority voting*) đối với bài toán phân lớp. Sự kết hợp này giúp làm giảm mức độ sai lệch của từng cây đơn lẻ và tạo ra một mô hình ổn định, đáng tin cậy hơn nhiều so với Decision Tree độc lập. Một trong những ưu điểm nổi bật của Random Forest là khả năng xử lý dữ liệu có số lượng thuộc tính lớn mà không yêu cầu giảm chiều trước. Thuật toán cũng có khả năng phát hiện các tương tác phi tuyến giữa các thuộc tính và tự động đánh giá mức độ quan trọng (feature importance) của từng biến trong quá trình phân loại.

Random Forest thường cho hiệu suất dự đoán cao ngay cả khi dữ liệu chứa nhiễu hoặc phân bố không đồng đều giữa các lớp. Điều này đặc biệt phù hợp với dataset của đồ án, nơi mức độ stress có sự phân bổ khác nhau giữa các mức, đồng thời các thuộc tính liên quan đến hành vi công nghệ, giấc ngủ và trạng thái tâm lý có thể chứa nhiều quan hệ phức tạp và có mức độ biến thiên lớn. Ngoài ra, Random Forest ít bị ảnh hưởng

bởi hiện tượng overfitting nhờ vào bản chất tổ hợp của mô hình, làm cho nó trở thành một lựa chọn tốt trong các bài toán dự đoán thực tế.

2.2.2 Lý do chọn thuật toán Random Forest

Việc lựa chọn Random Forest để giải quyết bài toán dự đoán mức độ căng thẳng trong đồ án xuất phát từ nhiều đặc điểm ưu việt của thuật toán so với các phương pháp phân lớp truyền thống. Random Forest là một mô hình tổ hợp mạnh, được xây dựng dựa trên việc kết hợp nhiều cây quyết định độc lập. Sự kết hợp này không chỉ làm tăng độ chính xác mà còn giúp mô hình ổn định và ít bị ảnh hưởng bởi nhiễu. Trong bối cảnh dữ liệu có sự đa dạng cao giữa các biến đầu vào và mối quan hệ phức tạp giữa các yếu tố tâm lý và hành vi sử dụng công nghệ, Random Forest thể hiện rõ ràng những lợi thế nổi bật.

Trước hết, Random Forest khắc phục được nhược điểm lớn nhất của Decision Tree là xu hướng bị overfitting khi mô hình quá phức tạp. Nhờ cơ chế lấy mẫu bootstrap và lựa chọn ngẫu nhiên một tập con thuộc tính tại mỗi nút phân chia, các cây trong rừng tạo ra sự đa dạng cần thiết, giúp mô hình tổng thể có khả năng tổng quát hóa tốt hơn. Điều này rất quan trọng đối với bộ dữ liệu của đồ án vốn có nhiều thuộc tính tác động qua lại, chẳng hạn như thời gian sử dụng mạng xã hội, mức độ lo âu và chất lượng giấc ngủ, những yếu tố có thể gây nhiễu hoặc không tuyến tính.

Bên cạnh đó, Random Forest hoạt động hiệu quả ngay cả khi dữ liệu chứa nhiều thuộc tính không thực sự quan trọng. Thuật toán tự động đánh giá và đo lường mức độ đóng góp của từng thuộc tính trong quá trình dự đoán thông qua tham số *feature importance*. Điều này mang lại giá trị lớn cho đồ án, giúp xác định những yếu tố then chốt ảnh hưởng đến mức độ căng thẳng của người dùng, chẳng hạn như số giờ sử dụng thiết bị, thời lượng ngủ hoặc mức độ lo âu. Việc phân tích này góp phần làm sáng tỏ mối quan hệ giữa công nghệ và sức khỏe tinh thần, phù hợp với mục tiêu nghiên cứu đã đặt ra.

Một lý do khác để lựa chọn Random Forest là khả năng xử lý tốt dữ liệu kích thước lớn cũng như dữ liệu có nhiều chiều. Thuật toán này không yêu cầu chuẩn hóa dữ liệu đầu vào và có thể tương tác tốt với cả biến định lượng và biến phân loại. Điều này giúp giảm thiểu khối lượng công việc tiền xử lý dữ liệu và đảm bảo rằng mô hình không bị ảnh hưởng quá mức bởi sự khác biệt trong thang đo của các thuộc tính.

Ngoài ra, Random Forest có ưu điểm nổi bật về độ ổn định của kết quả. Do mô hình dựa trên nhiều cây quyết định được xây dựng độc lập, sự thay đổi nhỏ trong dữ liệu không làm thay đổi đáng kể kết quả cuối cùng. Điều này khác biệt với Decision Tree, nơi chỉ một biến động nhỏ cũng có thể làm thay đổi đáng kể cấu trúc cây. Tính ổn định này đặc biệt hữu ích trong bài toán dự đoán stress, nơi dữ liệu thu thập có thể có sai số hoặc biến động nhẹ do yếu tố tâm lý và thói quen sinh hoạt vốn rất đa dạng.

Cuối cùng, Random Forest còn hỗ trợ song song hóa quá trình huấn luyện, giúp rút ngắn thời gian xử lý ngay cả khi số lượng cây lớn. Nhờ tốc độ huấn luyện tốt và hiệu suất dự đoán cao, Random Forest thường được xem là thuật toán mặc định mạnh mẽ cho nhiều bài toán phân lớp trong lĩnh vực khoa học dữ liệu.

Từ các phân tích trên, có thể thấy Random Forest là lựa chọn phù hợp cho bài toán dự đoán mức độ căng thẳng trong đồ án, bởi nó kết hợp được độ chính xác cao, khả năng tổng quát tốt, tính ổn định và khả năng giải thích mức độ quan trọng của các đặc trưng. Đây là những yếu tố quan trọng để xây dựng một mô hình đáng tin cậy phục vụ phân tích khoa học.

2.2.3 Tập huấn luyện (Train Set) của thuật toán Random Forest

Quá trình xây dựng mô hình Random Forest trong đồ án được thực hiện dựa trên tập huấn luyện được tách ra từ bộ dữ liệu ban đầu thông qua hàm *train_test_split* của thư viện Scikit-Learn. Tương tự như mô hình Decision Tree, dữ liệu được phân chia theo tỷ lệ 80% dành cho huấn luyện và 20% dành cho kiểm thử. Tập huấn luyện này bao gồm phần lớn các mẫu dữ liệu mô tả hành vi sử dụng công nghệ, các chỉ số tâm lý và lối sống của người dùng.

Trong giai đoạn huấn luyện, mô hình Random Forest tạo ra nhiều cây quyết định bằng cách áp dụng cơ chế bootstrap sampling, trong đó mỗi cây được huấn luyện từ một tập con dữ liệu được lấy mẫu ngẫu nhiên có hoàn lại từ train set. Điều này giúp mô hình xây dựng được nhiều cấu trúc cây khác nhau, thể hiện những khía cạnh khác nhau trong mối quan hệ giữa các thuộc tính đầu vào và biến mục tiêu *stress_level*.

Ngoài ra, Random Forest còn sử dụng cơ chế lựa chọn ngẫu nhiên một số lượng thuộc tính tại mỗi nút phân chia. Do đó, tập huấn luyện đóng vai trò quan trọng không chỉ trong việc hình thành các cây thành phần mà còn ảnh hưởng đến cách thuật toán học

được các quy tắc phân loại thông qua các tập thuộc tính khác nhau. Điều này giúp mô hình giảm thiểu tình trạng overfitting và tăng khả năng tổng quát hóa.

Trong chương trình của đồ án, train set được đưa vào mô hình Random Forest để thực hiện quá trình học bằng việc điều chỉnh tự động các cây phân loại, từ đó tạo nên một mô hình mạnh có khả năng dự đoán chính xác mức độ căng thẳng dựa trên các đặc trưng đầu vào. Train set do đó đóng vai trò nền tảng, quyết định chất lượng của mô hình Random Forest trong suốt quá trình thực nghiệm.

age	gender	daily_screen_time_hours	sleep_duration_hours	social_media_hours	work_related_hours	gaming_hours	phone_usage_hours
0.423728814	0	0.533333333	0.555555556	0.928571429	0.736842105	0.315789474	0.520833333
0.542372881	0	0.188888889	0.583333333	0.071428571	0.184210526	0.473684211	0.375
0.711864407	0.5	0.377777778	0.527777778	1	0.894736842	0.368421053	0.416666667
0.745762712	0	0.422222222	0.666666667	0.666666667	0.868421053	0.447368421	0.479166667
0.152542373	0	0.744444444	0.222222222	0.857142857	0.815789474	0.289473684	0.3125
0.372881356	0.5	0.477777778	0.611111111	0.880952381	0.815789474	0.157894737	0.5
0.610169492	0.5	0.522222222	0.527777778	1	0.894736842	0.342105263	0.770833333
0.576271186	0.5	0.011111111	0.666666667	0.880952381	0.815789474	0.289473684	0.291666667
0.237288136	0.5	0.588888889	0.666666667	0.833333333	0.921052632	0.210526316	0.458333333
0.813559322	1	0.633333333	0.361111111	0.952380952	0.815789474	0.421052632	0.583333333
0.847457627	0	0.711111111	0.277777778	1	0.842105263	0.210526316	0.208333333
0.610169492	1	0.077777778	0.555555556	0.428571429	0.552631579	0.526315789	0.1875
0.728813559	0	0.322222222	0.305555556	0.595238095	0.736842105	0.315789474	0.3125
0.525423729	0.5	0.511111111	0.5	0.476190476	0.736842105	0.368421053	0.520833333
0.610169492	0.5	0.533333333	0.611111111	1	0.921052632	0.394736842	0.395833333
0.050847458	0	0.544444444	0.472222222	1	0.947368421	0.263157895	0.3125
0.355932203	0.5	0.822222222	0.555555556	1	0.842105263	0.552631579	0.395833333
0.169491525	0.5	0.144444444	0.527777778	0.619047619	0.736842105	0.473684211	0.041666667
0.271186441	0	0.5	0.277777778	0.785714286	0.894736842	0.368421053	0.583333333
0.423728814	1	0.5	0.416666667	0.880952381	0.894736842	0.236842105	0.3125
0.338983051	0.5	0.133333333	0.833333333	0.142857143	0.210526316	0.710526316	0.541666667
0.762711864	0	0.255555556	0.527777778	0.976190476	0.947368421	0.421052632	0.291666667
0.559322034	0.5	0.266666667	0.5	0.595238095	0.605263158	0.447368421	0.520833333
0.644067797	0	0.277777778	0.555555556	0.476190476	0.605263158	0.421052632	0.375
0.457627119	1	0.6	0.277777778	1	0.815789474	0.210526316	0.583333333
0.898305085	1	0.444444444	0.638888889	0.80952381	0.815789474	0.342105263	0.25
0.898305085	0	0.455555556	0.472222222	0.952380952	0.842105263	0.210526316	0.270833333
0.06779661	0	0.322222222	0.888888889	0.761904762	0.815789474	0.447368421	0
0.525423729	0.5	0.155555556	0.777777778	0.523809524	0.710526316	0.736842105	0.0625
0.254237288	0.5	0.644444444	0.638888889	1	0.894736842	0.184210526	0.416666667
0.491525424	0.5	0.422222222	0.722222222	1	0.842105263	0.131578947	0.625
0.237288136	0.5	0.388888889	0.361111111	0.80952381	0.868421053	0.157894737	0.520833333
0.644067797	0.5	0.177777778	0.583333333	0.119047619	0.184210526	0.631578947	0.166666667
0.050847458	0	0.477777778	0.694444444	1	0.815789474	0.368421053	0.0625
1	0.5	0.711111111	0.444444444	0.785714286	0.710526316	0.184210526	0.479166667
0.186440678	0	0.7	0.694444444	1	0.868421053	0.263157895	0.479166667
0.440677966	0	0.477777778	0.583333333	1	0.894736842	0.105263158	0.729166667
0.423728814	0	0.544444444	0.416666667	0.5	0.631578947	0.710526316	0.208333333

Hình 2. 5 Tập data train random-forest 1

laptop_usage_hours	sleep_quality	health_score	sleep_health_index	emotional_balance	overall_wellness	digital_stress_score	work_life_balance
0.78	0.75	0.579710145	0.617021277	0.3	0.454545455	0.658862876	0.263157895
0.18	1	0.797101449	0.840425532	0.988888889	0.922459893	0.205685619	0.815789474
0.12	0.75	0.260869565	0.606382979	0.077777778	0.227272727	0.585284281	0.105263158
0.26	1	0.536231884	0.872340426	0.366666667	0.534759358	0.519230769	0.131578947
0.32	0.75	0.362318841	0.489361702	0.277777778	0.331550802	0.640468227	0.184210526
0.36	0.75	0.376811594	0.638297872	0.111111111	0.294117647	0.615384615	0.184210526
0.26	0.75	0.289855072	0.606382979	0.033333333	0.21657754	0.767558528	0.105263158
0.02	1	0.434782609	0.872340426	0.366666667	0.497326203	0.364548495	0.184210526
0	0.75	0.623188406	0.659574468	0.311111111	0.486631016	0.627090301	0.078947368
0.44	0.5	0.391304348	0.329787234	0	0.168449198	0.726588629	0.184210526
0.62	0.75	0.289855072	0.510638298	0.022222222	0.187165775	0.635451505	0.157894737
0	1	0.797101449	0.829787234	0.722222222	0.79144385	0.2090301	0.447368421
0.34	1	0.695652174	0.734042553	0.566666667	0.655080214	0.398829431	0.263157895
0.52	0.75	0.391304348	0.595744681	0.511111111	0.481283422	0.507525084	0.263157895
0.28	0.5	0.391304348	0.425531915	0.044444444	0.213903743	0.636287625	0.078947368
0.24	0.5	0.202898551	0.372340426	0	0.109625668	0.610367893	0.052631579
0.6	0.5	0.217391304	0.404255319	0	0.122994652	0.744983278	0.157894737
0.08	1	0.565217391	0.819148936	0.577777778	0.63368984	0.241638796	0.263157895
0.08	0.75	0.304347826	0.510638298	0.333333333	0.342245989	0.623745819	0.105263158
0.4	0.75	0.376811594	0.563829787	0.222222222	0.328877005	0.556020067	0.105263158
0.14	1	0.739130435	0.936170213	0.811111111	0.839572193	0.267558528	0.789473684
0.24	0.75	0.130434783	0.606382979	0.055555556	0.168449198	0.486622074	0.052631579
0.1	0.75	0.695652174	0.595744681	0.522222222	0.598930481	0.453177258	0.394736842
0	1	0.782608696	0.829787234	0.577777778	0.71657754	0.367056856	0.394736842
0.5	0.75	0.188405797	0.510638298	0	0.139037433	0.72909699	0.184210526
0.3	0.75	0.710144928	0.64893617	0.333333333	0.526737968	0.489966555	0.184210526
0.12	0.75	0.47826087	0.585106383	0.255555556	0.387700535	0.546822742	0.157894737
0.32	1	0.420289855	0.957446809	0.233333333	0.449197861	0.338628763	0.184210526
0.34	0.75	0.652173913	0.70212766	0.533333333	0.614973262	0.223244147	0.289473684
0.34	0.5	0.304347826	0.436170213	0	0.163101604	0.685618729	0.105263158
0.2	0.75	0.333333333	0.680851064	0	0.235294118	0.677257525	0.157894737
0	0.75	0.47826087	0.542553191	0.4	0.446524064	0.566889632	0.131578947
0.2	1	0.753623188	0.840425532	0.777777778	0.804812834	0.141304348	0.815789474
0.28	0.75	0.057971014	0.670212766	0	0.131016043	0.494983278	0.184210526
0.5	0.5	0.652173913	0.361702128	0.422222222	0.475935829	0.665551839	0.289473684
0.32	0.5	0.333333333	0.457446809	0.1	0.227272727	0.72909699	0.131578947
0.18	0.75	0.304347826	0.627659574	0.033333333	0.227272727	0.735785953	0.105263158
0.56	0.75	0.594202899	0.563829787	0.588888889	0.585561497	0.414715719	0.368421053

Hình 2. 6 Tập data test random-forest 2

2.2.4 Tập kiểm thử (Test Set) của thuật toán Random Forest

Tập kiểm thử của thuật toán Random Forest bao gồm 20% dữ liệu được tách ra ngay từ đầu và hoàn toàn không tham gia vào quá trình huấn luyện mô hình. Mục tiêu chính của test set là cung cấp một bộ dữ liệu độc lập để đánh giá khả năng tổng quát hóa của mô hình Random Forest sau khi đã được huấn luyện bằng train set.

Trong chương trình đồ án, sau khi mô hình hoàn thành quá trình huấn luyện, test set được đưa vào mô hình để dự đoán giá trị *stress_level* dựa trên các thuộc tính như thời gian sử dụng thiết bị, thời lượng ngủ, mức độ lo âu, trầm cảm và các đặc điểm hành vi khác. Kết quả dự đoán này sau đó được so sánh với giá trị *stress_level* thực tế trong test set để tính toán các chỉ số hiệu suất như độ chính xác (accuracy), ma trận nhầm lẫn (confusion matrix), độ nhạy (recall) và độ chính xác theo từng lớp (precision).

Việc đánh giá mô hình bằng test set mang ý nghĩa đặc biệt quan trọng đối với thuật toán Random Forest. Nhờ cấu trúc tổ hợp, Random Forest thường đạt hiệu suất dự đoán cao hơn Decision Tree đơn lẻ, và test set là công cụ giúp kiểm chứng điều này. Nếu mô hình duy trì được hiệu suất tốt trên test set, điều đó chứng tỏ Random Forest đã học được các quy luật tổng quát thay vì ghi nhớ dữ liệu huấn luyện. Ngược lại, nếu hiệu suất giảm sút đáng kể, cần xem xét lại số lượng cây, độ sâu tối đa hoặc các tham số điều chỉnh khác của mô hình.

Như vậy, test set đóng vai trò như thước đo cuối cùng để xác định chất lượng thực tế của mô hình và đảm bảo rằng Random Forest có khả năng dự đoán mức độ căng thẳng một cách chính xác và đáng tin cậy khi áp dụng vào dữ liệu mới.

age	gender	daily_screen_time_hours	sleep_duration_hours	social_media_hours	work_related_hours	gaming_hours	phone_usage_hours
0.576271186	0	0.577777778	0.416666667	1	0.921052632	0.473684211	0.395833333
0.949152542	0.5	0.344444444	0.972222222	0.547619048	0.657894737	0.605263158	0.229166667
0.898305085	0	0.644444444	0.222222222	0.80952381	0.815789474	0.184210526	0.645833333
0.525423729	0.5	0.711111111	0.611111111	1	0.815789474	0.263157895	0.645833333
0.542372881	0	0.433333333	0.638888889	0.404761905	0.421052632	0.605263158	0.3125
0.389830508	0	0.677777778	0.583333333	1	0.842105263	0.105263158	0.583333333
0.406779661	0.5	0.555555556	0.555555556	1	0.815789474	0.236842105	0.645833333
0.474576271	0.5	0.522222222	0.583333333	1	0.921052632	0.315789474	0.4375
0.847457627	0	0.588888889	0.416666667	0.904761905	0.815789474	0.289473684	0.375
0.559322034	0.5	0.411111111	0.277777778	0.642857143	0.684210526	0.447368421	0.375
0.593220339	0	0.577777778	0.666666667	1	0.921052632	0.473684211	0.291666667
0.779661017	0.5	0.333333333	0.694444444	0.357142857	0.526315789	0.578947368	0.1875
0.949152542	0	0.2	0.611111111	0.142857143	0.131578947	0.763157895	0.104166667
0.576271186	0.5	0.222222222	0.694444444	0.095238095	0.184210526	0.605263158	0.208333333
0.881355932	0.5	0.544444444	0.638888889	0.904761905	0.842105263	0.289473684	0.416666667
0.389830508	0.5	0.277777778	0.5	0.142857143	0.131578947	0.815789474	0.270833333
0.93220339	0	0.555555556	0.611111111	0.738095238	0.789473684	0.394736842	0.416666667
0.288135593	0	0.377777778	0.277777778	0.904761905	0.842105263	0.263157895	0.229166667
0.491525424	0.5	0.733333333	0.472222222	0.80952381	0.763157895	0.263157895	0.395833333
0.271186441	0	0.422222222	0.694444444	0.833333333	0.815789474	0.315789474	0.4375
0.050847458	0	0.944444444	0.5	1	0.921052632	0.184210526	0.708333333
0.86440678	0	0.344444444	0.611111111	0.547619048	0.763157895	0.473684211	0.416666667
0.406779661	0.5	0.322222222	0.555555556	0.761904762	0.815789474	0.421052632	0.354166667
0.508474576	0.5	0.577777778	0.25	1	0.789473684	0.421052632	0.666666667
0.610169492	0	0.233333333	0.694444444	0.142857143	0.052631579	0.815789474	0.3125
0.220338983	1	0.266666667	0.638888889	0.357142857	0.368421053	0.684210526	0.416666667
0.101694915	0.5	0.633333333	0.555555556	1	0.894736842	0.157894737	0.291666667
0.644067797	0.5	0.288888889	0.555555556	0.714285714	0.789473684	0.368421053	0.4375
0.457627119	0	0.733333333	0.75	1	0.815789474	0.078947368	0.479166667
0.610169492	0.5	0.544444444	0.583333333	0.80952381	0.789473684	0.368421053	0.604166667
0.440677966	0.5	0	0.694444444	0.857142857	0.868421053	0.447368421	0.125
0.983050847	0	0.588888889	0.555555556	0.333333333	0.473684211	0.605263158	0.416666667
0.186440678	0.5	0.566666667	0.416666667	1	0.789473684	0.289473684	0.375
0.254237288	0	0.333333333	0.472222222	0.714285714	0.815789474	0.289473684	0.520833333
0.305084746	0.5	0.333333333	0.527777778	0.666666667	0.736842105	0.552631579	0.4375
0.355932203	0	0.3	0.5	0.666666667	0.736842105	0.394736842	0.666666667
0.389830508	0	0.4	0.527777778	1	0.842105263	0.263157895	0.145833333
0.644067797	0	0.288888889	0.305555556	0.761904762	0.842105263	0.394736842	0.125

Hình 2. 7 Tập data train random-forest 1

laptop_usage_hours	sleep_quality	health_score	sleep_health_index	emotional_balance	overall_wellness	digital_stress_score	work_life_balance
0.46	0.5	0.347826087	0.35106383	0	0.157754011	0.653010033	0.078947368
0.24	1	0.550724638	0.989361702	0.677777778	0.719251337	0.362040134	0.342105263
0.36	0.75	0.565217391	0.489361702	0.3	0.417112299	0.70819398	0.184210526
0.24	0.75	0.449275362	0.638297872	0.177777778	0.352941176	0.793478261	0.184210526
0.14	0.75	0.724637681	0.64893617	0.722222222	0.719251337	0.380434783	0.578947368
0.54	0.5	0.173913043	0.414893617	0	0.109625668	0.758361204	0.157894737
0.58	0.75	0.391304348	0.617021277	0.022222222	0.251336898	0.734949833	0.184210526
0.66	0.75	0.304347826	0.627659574	0	0.211229947	0.647157191	0.078947368
0.38	0.75	0.47826087	0.563829787	0.288888889	0.398395722	0.619565217	0.184210526
0.1	1	0.666666667	0.723404255	0.6	0.657754011	0.469899666	0.315789474
0.24	1	0.434782609	0.872340426	0.111111111	0.374331551	0.615384615	0.078947368
0.5	0.75	0.608695652	0.670212766	0.8	0.719251337	0.282608696	0.473684211
0.34	1	0.710144928	0.85106383	0.844444444	0.823529412	0.134615385	0.868421053
0.04	1	0.797101449	0.882978723	0.9	0.890374332	0.165551839	0.815789474
0.3	0.75	0.362318841	0.64893617	0.288888889	0.377005348	0.617892977	0.157894737
0.36	0.75	0.666666667	0.595744681	1	0.818181818	0.224080268	0.868421053
0.46	1	0.724637681	0.85106383	0.511111111	0.668449198	0.569397993	0.210526316
0.68	0.75	0.347826087	0.510638298	0.177777778	0.28342246	0.487458194	0.157894737
0.44	0.75	0.420289855	0.585106383	0.311111111	0.393048128	0.651337793	0.236842105
0.36	0.75	0.492753623	0.670212766	0.155555556	0.36631016	0.556856187	0.184210526
0.46	0.5	0.31884058	0.382978723	0	0.155080214	0.903846154	0.078947368
0.62	0.75	0.52173913	0.638297872	0.5	0.534759358	0.429765886	0.236842105
0.38	1	0.608695652	0.829787234	0.455555556	0.593582888	0.466555184	0.184210526
0.22	0.5	0.260869565	0.287234043	0	0.109625668	0.75083612	0.210526316
0.3	1	0.898550725	0.882978723	1	0.975935829	0.222408027	0.947368421
0.22	0.75	0.797101449	0.64893617	0.855555556	0.810160428	0.340301003	0.631578947
0.12	0.75	0.188405797	0.617021277	0	0.165775401	0.636287625	0.105263158
0.14	0.75	0.463768116	0.617021277	0.422222222	0.470588235	0.469063545	0.210526316
0.64	0.75	0.31884058	0.691489362	0.066666667	0.264705882	0.741638796	0.184210526
0.46	0.75	0.463768116	0.627659574	0.144444444	0.339572193	0.655518395	0.210526316
0.02	0.75	0.507246377	0.670212766	0.233333333	0.409090909	0.29264214	0.131578947
0.6	0.75	0.695652174	0.617021277	0.788888889	0.732620321	0.454013378	0.526315789
0.42	0.75	0.405797101	0.563829787	0.077777778	0.270053476	0.641304348	0.210526316
0.28	1	0.565217391	0.79787234	0.444444444	0.564171123	0.515886288	0.184210526
0.26	0.75	0.623188406	0.606382979	0.455555556	0.542780749	0.470735786	0.263157895
0.16	0.75	0.608695652	0.595744681	0.533333333	0.572192513	0.5409699	0.263157895
0.52	0.75	0.173913043	0.606382979	0.255555556	0.280748663	0.495819398	0.157894737
0.64	0.75	0.362318841	0.521276596	0.444444444	0.419786096	0.371237458	0.157894737

Hình 2. 8 Tập data test random-forest 2

2.3 Sử dụng thuật toán K-Means

Trong đồ án, bên cạnh các mô hình phân lớp như Decision Tree và Random Forest được sử dụng để dự đoán mức độ căng thẳng của người dùng, thuật toán phân cụm K-Means được áp dụng nhằm phân chia người dùng thành các nhóm hành vi và trạng thái tâm lý tương đồng. Việc kết hợp giữa mô hình phân lớp và mô hình phân cụm giúp mở rộng khả năng phân tích dữ liệu theo hướng khám phá và nhóm hóa, từ đó mang lại cái nhìn đa chiều hơn về tác động của công nghệ đối với sức khỏe tinh thần.

Thuật toán K-Means hoạt động theo phương pháp học không giám sát, nghĩa là không sử dụng biến mục tiêu để học mô hình. Thay vào đó, thuật toán tìm kiếm cấu trúc tiềm ẩn trong dữ liệu bằng cách phân chia các quan sát vào những nhóm (clusters) sao cho các điểm dữ liệu trong cùng một nhóm có mức độ tương đồng cao, còn các điểm thuộc nhóm khác thì khác biệt nhiều nhất có thể. Trong bối cảnh của đồ án, K-Means

được sử dụng để phát hiện các nhóm người dùng có hành vi sử dụng thiết bị công nghệ tương tự nhau, ví dụ nhóm sử dụng mạng xã hội nhiều, nhóm tập trung vào công việc, nhóm có thói quen sử dụng cân bằng, hoặc nhóm có biểu hiện liên quan đến stress cao.

Kết quả phân cụm giúp hỗ trợ cho việc phân tích nguyên nhân dẫn đến stress ở từng nhóm người dùng, đồng thời đóng vai trò quan trọng trong việc đề xuất giải pháp cải thiện sức khỏe tinh thần mang tính cá nhân hóa. Việc áp dụng thuật toán K-Means vì thế bổ sung cho các mô hình dự đoán, tạo nên hệ thống phân tích toàn diện hơn.

2.3.1 Tổng quan về thuật toán phân cụm dựa trên K-Means

K-Means là một trong những thuật toán phân cụm phổ biến nhất trong khai phá dữ liệu, được sử dụng rộng rãi nhờ tính đơn giản, tốc độ nhanh và khả năng xử lý hiệu quả các bộ dữ liệu lớn. Thuật toán hoạt động dựa trên nguyên tắc tối thiểu hóa tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm (centroid) của chúng. Mỗi cụm được biểu diễn bằng một tâm cụm, là giá trị trung bình của các điểm thuộc cụm đó.

K-Means bắt đầu bằng việc lựa chọn số lượng cụm k, sau đó khởi tạo ngẫu nhiên các tâm cụm ban đầu. Quá trình phân cụm tiếp tục lặp lại hai bước chính: (1) gán mỗi điểm dữ liệu vào cụm có tâm gần nhất dựa trên độ đo khoảng cách, thường là khoảng cách Euclid; (2) cập nhật tâm cụm bằng cách tính trung bình các điểm đã được gán vào mỗi cụm. Thuật toán lặp lại hai bước này cho đến khi các tâm cụm hội tụ hoặc khi thay đổi giữa hai vòng lặp liên tiếp không đáng kể.

Trong bối cảnh phân tích dữ liệu hành vi, K-Means có khả năng phát hiện các nhóm người dùng dựa trên những đặc điểm tương đồng về thời gian sử dụng thiết bị, mức độ tham gia mạng xã hội, thời gian làm việc trên máy tính, thời lượng ngủ hay các thông số về sức khỏe tinh thần. Vì không yêu cầu biến mục tiêu, thuật toán giúp khai phá các mẫu hành vi tiềm ẩn mà không bị ràng buộc bởi nhãn dữ liệu.

Điểm mạnh của K-Means nằm ở khả năng xử lý dữ liệu có số lượng lớn thuộc tính dạng số, đồng thời tạo ra phân cụm dễ diễn giải. Tuy nhiên, thuật toán cũng nhạy cảm với dữ liệu nhiễu và yêu cầu chuẩn hóa dữ liệu để đảm bảo các thuộc tính có thang đo khác nhau không gây ảnh hưởng đến khoảng cách đo lường. Đối với bộ dữ liệu của đồ án, các đặc trưng như thời gian sử dụng thiết bị, số giờ hoạt động, điểm lo âu hoặc trầm cảm đều ở dạng số, rất phù hợp với yêu cầu của thuật toán K-Means.

Nhìn chung, K-Means đóng vai trò quan trọng trong việc phân tích nhóm đối tượng theo hành vi sử dụng công nghệ và các yếu tố tâm lý. Kết quả phân cụm cung cấp thông tin bổ sung cho mô hình dự đoán stress, giúp xây dựng một bức tranh toàn diện và sâu sắc hơn về ảnh hưởng của công nghệ đến sức khỏe tinh thần của người dùng.

2.3.2 Lý do chọn thuật toán K-Means

Việc lựa chọn thuật toán K-Means trong đồ án xuất phát từ mục tiêu phân tích hành vi người dùng theo hướng khám phá các cấu trúc tiềm ẩn trong dữ liệu. Khác với Decision Tree hay Random Forest – những thuật toán học có giám sát tập trung vào nhiệm vụ dự đoán mức độ căng thẳng – K-Means được sử dụng để phân cụm nhằm nhận diện các nhóm người dùng có đặc điểm tương đồng, từ đó mang lại góc nhìn bổ sung trong việc đánh giá tác động của công nghệ đối với sức khỏe tinh thần.

Một trong những lý do quan trọng để lựa chọn K-Means là tính đơn giản, hiệu quả và khả năng xử lý các bộ dữ liệu lớn với chi phí tính toán thấp. Bộ dữ liệu trong đồ án gồm 5000 quan sát và nhiều thuộc tính có bản chất định lượng, chẳng hạn như thời gian sử dụng thiết bị, thời lượng ngủ, điểm lo âu và trầm cảm. Các thuộc tính này phù hợp với điều kiện áp dụng của K-Means, vốn hoạt động tốt khi xử lý dữ liệu dạng số và đo lường khoảng cách giữa các quan sát.

Bên cạnh đó, K-Means có khả năng phát hiện các nhóm hành vi một cách tự nhiên dựa trên sự tương đồng về khoảng cách trong không gian đặc trưng. Điều này rất quan trọng đối với bài toán của đồ án, bởi các nhóm người dùng có thể biểu hiện những mẫu hành vi công nghệ khác nhau, ví dụ như nhóm sử dụng mạng xã hội nhiều, nhóm làm việc với máy tính trong thời gian dài, nhóm ngủ ít, hoặc nhóm có điểm lo âu và trầm cảm cao. Việc phân chia thành các nhóm như vậy giúp hiểu sâu hơn về những yếu tố gây stress trong từng nhóm người dùng cụ thể và có thể hỗ trợ xây dựng các khuyến nghị phù hợp.

Một lý do khác khiến K-Means được lựa chọn là khả năng trực quan hóa kết quả phân cụm một cách dễ dàng. Các tâm cụm (centroids) mà thuật toán tạo ra cho phép mô tả đặc trưng trung bình của từng nhóm, giúp đánh giá xem nhóm nào có mức sử dụng công nghệ cao, nhóm nào có lối sống lành mạnh hay nhóm nào có nguy cơ stress cao. Việc giải thích kết quả từ K-Means vì thế tương đối trực quan, hỗ trợ tốt cho quá trình báo cáo và trình bày kết quả nghiên cứu.

K-Means cũng có tính linh hoạt cao nhờ khả năng điều chỉnh số lượng cụm (k). Điều này cho phép người nghiên cứu thử nghiệm nhiều kịch bản khác nhau để tìm ra số cụm phù hợp nhất với cấu trúc dữ liệu. Bộ dữ liệu trong đồ án có độ đa dạng lớn, do đó việc lựa chọn k tối ưu giúp mô hình phân cụm phát hiện chính xác các nhóm hành vi đặc trưng.

Ngoài ra, thuật toán K-Means đã được tích hợp tối ưu trong các thư viện khoa học dữ liệu như Scikit-Learn, giúp quá trình triển khai trở nên đơn giản, nhanh chóng và dễ dàng tái lập. Với hiệu suất xử lý cao, K-Means phù hợp để áp dụng trong quá trình thực nghiệm của đồ án, giúp kiểm chứng các giả thuyết về nhóm hành vi người dùng và mối liên hệ giữa công nghệ và trạng thái tâm lý.

Tóm lại, K-Means được lựa chọn trong đồ án vì khả năng phân cụm hiệu quả, dễ diễn giải, phù hợp với dữ liệu dạng số và có khả năng hỗ trợ mạnh mẽ cho mục tiêu phân tích hành vi người dùng. Sự kết hợp giữa phân cụm và mô hình phân lớp contribuir mang lại cái nhìn toàn diện hơn về tác động của công nghệ đối với sức khỏe tinh thần của con người.

2.3.3 Tập huấn luyện (Train Set) của thuật toán K-Means

Trong mô hình phân cụm K-Means, vì không tồn tại biến mục tiêu để huấn luyện theo cách có giám sát, nên toàn bộ dữ liệu sau quá trình tiền xử lý được sử dụng để xây dựng mô hình phân cụm. Trong chương trình của đồ án, không có thao tác tách dữ liệu thành tập huấn luyện và tập kiểm thử như các thuật toán phân lớp. Thay vào đó, mô hình sử dụng toàn bộ các quan sát nhằm xác định cấu trúc nhóm tiềm ẩn trong dữ liệu dựa trên sự tương đồng về các đặc trưng hành vi và trạng thái tâm lý của người dùng.

Trước khi đưa vào mô hình K-Means, dữ liệu được chuẩn hóa bằng bộ *StandardScaler* đã được lưu sẵn trong chương trình. Chuẩn hóa này giúp đưa các thuộc tính về cùng thang đo, đảm bảo rằng những biến có giá trị lớn như thời gian sử dụng thiết bị hoặc lượng caffeine không lấn át các biến có giá trị nhỏ hơn như điểm chất lượng giấc ngủ. Điều này đặc biệt quan trọng đối với K-Means, vì thuật toán sử dụng khoảng cách Euclid làm tiêu chí phân cụm và rất nhạy cảm với sự khác biệt thang đo giữa các thuộc tính.

Trong giai đoạn huấn luyện, K-Means cố gắng phân chia dữ liệu thành k nhóm sao cho khoảng cách trong cụm là nhỏ nhất, còn khoảng cách giữa các cụm là lớn nhất.

Các tâm cụm (centroids) được cập nhật liên tục qua các vòng lặp cho đến khi hội tụ. Quá trình này không cần nhãn dữ liệu mà chỉ phụ thuộc vào cấu trúc phân bố tự nhiên của các điểm dữ liệu trong không gian đặc trưng.

Như vậy, "train set" trong thuật ngữ của mô hình học không giám sát thực chất là **toàn bộ tập dữ liệu sau chuẩn hóa**, được sử dụng để huấn luyện mô hình phân cụm K-Means và xác định tâm các cụm đặc trưng.

age	gender	daily_screen_time_hours	sleep_duration_hours	social_media_hours	work_related_hours	gaming_hours	phone_usage_hours
0.423728814	0	0.533333333	0.555555556	0.928571429	0.736842105	0.315789474	0.520833333
0.542372881	0	0.188888889	0.583333333	0.071428571	0.184210526	0.473684211	0.375
0.711864407	0.5	0.377777778	0.527777778	1	0.894736842	0.368421053	0.416666667
0.745762712	0	0.422222222	0.666666667	0.666666667	0.868421053	0.447368421	0.479166667
0.152542373	0	0.744444444	0.222222222	0.857142857	0.815789474	0.289473684	0.3125
0.372881356	0.5	0.477777778	0.611111111	0.880952381	0.815789474	0.157894737	0.5
0.610169492	0.5	0.522222222	0.527777778	1	0.894736842	0.342105263	0.770833333
0.576271186	0.5	0.011111111	0.666666667	0.880952381	0.815789474	0.289473684	0.291666667
0.237288136	0.5	0.588888889	0.666666667	0.833333333	0.921052632	0.210526316	0.458333333
0.813559322	1	0.633333333	0.361111111	0.952380952	0.815789474	0.421052632	0.583333333
0.847457627	0	0.711111111	0.277777778	1	0.842105263	0.210526316	0.208333333
0.610169492	1	0.077777778	0.555555556	0.428571429	0.552631579	0.526315789	0.1875
0.728813559	0	0.322222222	0.305555556	0.595238095	0.736842105	0.315789474	0.3125
0.525423729	0.5	0.511111111	0.5	0.476190476	0.736842105	0.368421053	0.520833333
0.610169492	0.5	0.533333333	0.611111111	1	0.921052632	0.394736842	0.395833333
0.050847458	0	0.544444444	0.472222222	1	0.947368421	0.263157895	0.3125
0.355932203	0.5	0.822222222	0.555555556	1	0.842105263	0.552631579	0.395833333
0.169491525	0.5	0.144444444	0.527777778	0.619047619	0.736842105	0.473684211	0.041666667
0.271186441	0	0.5	0.277777778	0.785714286	0.894736842	0.368421053	0.583333333
0.423728814	1	0.5	0.416666667	0.880952381	0.894736842	0.236842105	0.3125
0.338983051	0.5	0.133333333	0.833333333	0.142857143	0.210526316	0.710526316	0.541666667
0.762711864	0	0.255555556	0.527777778	0.976190476	0.947368421	0.421052632	0.291666667
0.559322034	0.5	0.266666667	0.5	0.595238095	0.605263158	0.447368421	0.520833333
0.644067797	0	0.277777778	0.555555556	0.476190476	0.605263158	0.421052632	0.375
0.457627119	1	0.6	0.277777778	1	0.815789474	0.210526316	0.583333333
0.898305085	1	0.444444444	0.638888889	0.80952381	0.815789474	0.342105263	0.25
0.898305085	0	0.455555556	0.472222222	0.952380952	0.842105263	0.210526316	0.270833333
0.06779661	0	0.322222222	0.888888889	0.761904762	0.815789474	0.447368421	0
0.525423729	0.5	0.155555556	0.777777778	0.523809524	0.710526316	0.736842105	0.0625
0.254237288	0.5	0.644444444	0.638888889	1	0.894736842	0.184210526	0.416666667
0.491525424	0.5	0.422222222	0.722222222	1	0.842105263	0.131578947	0.625
0.237288136	0.5	0.388888889	0.361111111	0.80952381	0.868421053	0.157894737	0.520833333
0.644067797	0.5	0.177777778	0.583333333	0.119047619	0.184210526	0.631578947	0.166666667
0.050847458	0	0.477777778	0.694444444	1	0.815789474	0.368421053	0.0625
1	0.5	0.711111111	0.444444444	0.785714286	0.710526316	0.184210526	0.479166667
0.186440678	0	0.7	0.694444444	1	0.868421053	0.263157895	0.479166667
0.440677966	0	0.477777778	0.583333333	1	0.894736842	0.105263158	0.729166667
0.423728814	0	0.544444444	0.416666667	0.5	0.631578947	0.710526316	0.208333333

Hình 2. 9 Tập data train k-means 1

laptop_usage_hours	sleep_quality	health_score	sleep_health_index	emotional_balance	overall_wellness	digital_stress_score	work_life_balance
0.78	0.75	0.579710145	0.617021277	0.3	0.454545455	0.658862876	0.263157895
0.18	1	0.797101449	0.840425532	0.988888889	0.922459893	0.205685619	0.815789474
0.12	0.75	0.260869565	0.606382979	0.077777778	0.227272727	0.585284281	0.105263158
0.26	1	0.536231884	0.872340426	0.366666667	0.534759358	0.519230769	0.131578947
0.32	0.75	0.362318841	0.489361702	0.277777778	0.331550802	0.640468227	0.184210526
0.36	0.75	0.376811594	0.638297872	0.111111111	0.294117647	0.615384615	0.184210526
0.26	0.75	0.289855072	0.606382979	0.033333333	0.21657754	0.767558528	0.105263158
0.02	1	0.434782609	0.872340426	0.366666667	0.497326203	0.364548495	0.184210526
0	0.75	0.623188406	0.659574468	0.311111111	0.486631016	0.627090301	0.078947368
0.44	0.5	0.391304348	0.329787234	0	0.168449198	0.726588629	0.184210526
0.62	0.75	0.289855072	0.510638298	0.022222222	0.187165775	0.635451505	0.157894737
0	1	0.797101449	0.829787234	0.722222222	0.79144385	0.2090301	0.447368421
0.34	1	0.695652174	0.734042553	0.566666667	0.655080214	0.398829431	0.263157895
0.52	0.75	0.391304348	0.595744681	0.511111111	0.481283422	0.507525084	0.263157895
0.28	0.5	0.391304348	0.425531915	0.044444444	0.213903743	0.636287625	0.078947368
0.24	0.5	0.202898551	0.372340426	0	0.109625668	0.610367893	0.052631579
0.6	0.5	0.217391304	0.404255319	0	0.122994652	0.744983278	0.157894737
0.08	1	0.565217391	0.819148936	0.577777778	0.63368984	0.241638796	0.263157895
0.08	0.75	0.304347826	0.510638298	0.333333333	0.342245989	0.623745819	0.105263158
0.4	0.75	0.376811594	0.563829787	0.222222222	0.328877005	0.556020067	0.105263158
0.14	1	0.739130435	0.936170213	0.811111111	0.839572193	0.267558528	0.789473684
0.24	0.75	0.130434783	0.606382979	0.055555556	0.168449198	0.486622074	0.052631579
0.1	0.75	0.695652174	0.595744681	0.522222222	0.598930481	0.453177258	0.394736842
0	1	0.782608696	0.829787234	0.577777778	0.71657754	0.367056856	0.394736842
0.5	0.75	0.188405797	0.510638298	0	0.139037433	0.72909699	0.184210526
0.3	0.75	0.710144928	0.64893617	0.333333333	0.526737968	0.489966555	0.184210526
0.12	0.75	0.47826087	0.585106383	0.255555556	0.387700535	0.546822742	0.157894737
0.32	1	0.420289855	0.957446809	0.233333333	0.449197861	0.338628763	0.184210526
0.34	0.75	0.652173913	0.70212766	0.533333333	0.614973262	0.223244147	0.289473684
0.34	0.5	0.304347826	0.436170213	0	0.163101604	0.685618729	0.105263158
0.2	0.75	0.333333333	0.680851064	0	0.235294118	0.677257525	0.157894737
0	0.75	0.47826087	0.542553191	0.4	0.446524064	0.566889632	0.131578947
0.2	1	0.753623188	0.840425532	0.777777778	0.804812834	0.141304348	0.815789474
0.28	0.75	0.057971014	0.670212766	0	0.131016043	0.494983278	0.184210526
0.5	0.5	0.652173913	0.361702128	0.422222222	0.475935829	0.665551839	0.289473684
0.32	0.5	0.333333333	0.457446809	0.1	0.227272727	0.72909699	0.131578947
0.18	0.75	0.304347826	0.627659574	0.033333333	0.227272727	0.735785953	0.105263158
0.56	0.75	0.594202899	0.563829787	0.588888889	0.585561497	0.414715719	0.368421053

Hình 2. 10 Tập data train k-means 2

2.3.4 Tập kiểm thử (Test Set) của thuật toán K-Means

Do tính chất không giám sát, K-Means không sử dụng một "test set" theo nghĩa truyền thống để đánh giá độ chính xác dự đoán như trong Decision Tree hoặc Random Forest. Tuy nhiên, chương trình của đồ án vẫn thực hiện bước đánh giá mô hình phân cụm bằng cách **áp dụng lại mô hình K-Means lên toàn bộ tập dữ liệu sau khi phân cụm** và xem xét mức độ hợp lý của các cụm được tạo ra.

Trong chương trình, quá trình đánh giá được thực hiện thông qua việc phân tích đặc trưng của từng cụm người dùng. Mỗi quan sát sau khi được chuẩn hóa sẽ được gán vào cụm gần nhất dựa trên khoảng cách đến tâm cụm. Khi các cụm đã được hình thành, người nghiên cứu tiến hành quan sát các chỉ số đặc trưng của từng cụm, chẳng hạn như:

- mức độ sử dụng mạng xã hội trung bình,

- thời gian làm việc với công nghệ,
- trạng thái tâm lý (lo âu, trầm cảm),
- chất lượng giấc ngủ,
- các đặc điểm hành vi nổi bật khác.

Thông qua việc phân tích này, mô hình K-Means được đánh giá dựa trên tính hợp lý và khả năng phân biệt giữa các cụm. Nếu các cụm có sự khác biệt rõ ràng về đặc trưng hành vi và tâm lý, điều đó chứng tỏ mô hình phân cụm hoạt động tốt. Ngược lại, nếu các cụm chồng chéo nhau hoặc không tạo ra sự khác biệt có ý nghĩa thống kê, mô hình có thể cần điều chỉnh lại số lượng cụm hoặc tham số thuật toán.

age	gender	daily_screen_time_hours	sleep_duration_hours	social_media_hours	work_related_hours	gaming_hours	phone_usage_hours
0.576271186	0	0.577777778	0.416666667	1	0.921052632	0.473684211	0.395833333
0.949152542	0.5	0.344444444	0.972222222	0.547619048	0.657894737	0.605263158	0.229166667
0.898305085	0	0.644444444	0.222222222	0.80952381	0.815789474	0.184210526	0.645833333
0.525423729	0.5	0.711111111	0.611111111	1	0.815789474	0.263157895	0.645833333
0.542372881	0	0.433333333	0.638888889	0.404761905	0.421052632	0.605263158	0.3125
0.389830508	0	0.677777778	0.583333333	1	0.842105263	0.105263158	0.583333333
0.406779661	0.5	0.555555556	0.555555556	1	0.815789474	0.236842105	0.645833333
0.474576271	0.5	0.522222222	0.583333333	1	0.921052632	0.315789474	0.4375
0.847457627	0	0.588888889	0.416666667	0.904761905	0.815789474	0.289473684	0.375
0.559322034	0.5	0.411111111	0.277777778	0.642857143	0.684210526	0.447368421	0.375
0.593220339	0	0.577777778	0.666666667	1	0.921052632	0.473684211	0.291666667
0.779661017	0.5	0.333333333	0.694444444	0.357142857	0.526315789	0.578947368	0.1875
0.949152542	0	0.2	0.611111111	0.142857143	0.131578947	0.763157895	0.104166667
0.576271186	0.5	0.222222222	0.694444444	0.095238095	0.184210526	0.605263158	0.208333333
0.881355932	0.5	0.544444444	0.638888889	0.904761905	0.842105263	0.289473684	0.416666667
0.389830508	0.5	0.277777778	0.5	0.142857143	0.131578947	0.815789474	0.270833333
0.93220339	0	0.555555556	0.611111111	0.738095238	0.789473684	0.394736842	0.416666667
0.288135593	0	0.377777778	0.277777778	0.904761905	0.842105263	0.263157895	0.229166667
0.491525424	0.5	0.733333333	0.472222222	0.80952381	0.763157895	0.263157895	0.395833333
0.271186441	0	0.422222222	0.694444444	0.833333333	0.815789474	0.315789474	0.4375
0.050847458	0	0.944444444	0.5	1	0.921052632	0.184210526	0.708333333
0.86440678	0	0.344444444	0.611111111	0.547619048	0.763157895	0.473684211	0.416666667
0.406779661	0.5	0.322222222	0.555555556	0.761904762	0.815789474	0.421052632	0.354166667
0.508474576	0.5	0.577777778	0.25	1	0.789473684	0.421052632	0.666666667
0.610169492	0	0.233333333	0.694444444	0.142857143	0.052631579	0.815789474	0.3125
0.220338983	1	0.266666667	0.638888889	0.357142857	0.368421053	0.684210526	0.416666667
0.101694915	0.5	0.633333333	0.555555556	1	0.894736842	0.157894737	0.291666667
0.644067797	0.5	0.288888889	0.555555556	0.714285714	0.789473684	0.368421053	0.4375
0.457627119	0	0.733333333	0.75	1	0.815789474	0.078947368	0.479166667
0.610169492	0.5	0.544444444	0.583333333	0.80952381	0.789473684	0.368421053	0.604166667
0.440677966	0.5	0	0.694444444	0.857142857	0.868421053	0.447368421	0.125
0.983050847	0	0.588888889	0.555555556	0.333333333	0.473684211	0.605263158	0.416666667
0.186440678	0.5	0.566666667	0.416666667	1	0.789473684	0.289473684	0.375
0.254237288	0	0.333333333	0.472222222	0.714285714	0.815789474	0.289473684	0.520833333
0.305084746	0.5	0.333333333	0.527777778	0.666666667	0.736842105	0.552631579	0.4375
0.355932203	0	0.3	0.5	0.666666667	0.736842105	0.394736842	0.666666667
0.389830508	0	0.4	0.527777778	1	0.842105263	0.263157895	0.145833333
0.644067797	0	0.288888889	0.305555556	0.761904762	0.842105263	0.394736842	0.125

Hình 2. 11 Tập data test k-means 1

laptop_usage_hours	sleep_quality	health_score	sleep_health_index	emotional_balance	overall_wellness	digital_stress_score	work_life_balance
0.46	0.5	0.347826087	0.35106383	0	0.157754011	0.653010033	0.078947368
0.24	1	0.550724638	0.989361702	0.677777778	0.719251337	0.362040134	0.342105263
0.36	0.75	0.565217391	0.489361702	0.3	0.417112299	0.70819398	0.184210526
0.24	0.75	0.449275362	0.638297872	0.177777778	0.352941176	0.793478261	0.184210526
0.14	0.75	0.724637681	0.64893617	0.722222222	0.719251337	0.380434783	0.578947368
0.54	0.5	0.173913043	0.414893617	0	0.109625668	0.758361204	0.157894737
0.58	0.75	0.391304348	0.617021277	0.022222222	0.251336898	0.734949833	0.184210526
0.66	0.75	0.304347826	0.627659574	0	0.211229947	0.647157191	0.078947368
0.38	0.75	0.47826087	0.563829787	0.288888889	0.398395722	0.619565217	0.184210526
0.1	1	0.666666667	0.723404255	0.6	0.657754011	0.469899666	0.315789474
0.24	1	0.434782609	0.872340426	0.111111111	0.374331551	0.615384615	0.078947368
0.5	0.75	0.608695652	0.670212766	0.8	0.719251337	0.282608696	0.473684211
0.34	1	0.710144928	0.85106383	0.844444444	0.823529412	0.134615385	0.868421053
0.04	1	0.797101449	0.882978723	0.9	0.890374332	0.165551839	0.815789474
0.3	0.75	0.362318841	0.64893617	0.288888889	0.377005348	0.617892977	0.157894737
0.36	0.75	0.666666667	0.595744681	1	0.818181818	0.224080268	0.868421053
0.46	1	0.724637681	0.85106383	0.511111111	0.668449198	0.569397993	0.210526316
0.68	0.75	0.347826087	0.510638298	0.177777778	0.28342246	0.487458194	0.157894737
0.44	0.75	0.420289855	0.585106383	0.311111111	0.393048128	0.651337793	0.236842105
0.36	0.75	0.492753623	0.670212766	0.155555556	0.36631016	0.556856187	0.184210526
0.46	0.5	0.31884058	0.382978723	0	0.155080214	0.903846154	0.078947368
0.62	0.75	0.52173913	0.638297872	0.5	0.534759358	0.429765886	0.236842105
0.38	1	0.608695652	0.829787234	0.455555556	0.593582888	0.466555184	0.184210526
0.22	0.5	0.260869565	0.287234043	0	0.109625668	0.75083612	0.210526316
0.3	1	0.898550725	0.882978723	1	0.975935829	0.222408027	0.947368421
0.22	0.75	0.797101449	0.64893617	0.855555556	0.810160428	0.340301003	0.631578947
0.12	0.75	0.188405797	0.617021277	0	0.165775401	0.636287625	0.105263158
0.14	0.75	0.463768116	0.617021277	0.422222222	0.470588235	0.469063545	0.210526316
0.64	0.75	0.31884058	0.691489362	0.066666667	0.264705882	0.741638796	0.184210526
0.46	0.75	0.463768116	0.627659574	0.144444444	0.339572193	0.655518395	0.210526316
0.02	0.75	0.507246377	0.670212766	0.233333333	0.409090909	0.29264214	0.131578947
0.6	0.75	0.695652174	0.617021277	0.788888889	0.732620321	0.454013378	0.526315789
0.42	0.75	0.405797101	0.563829787	0.077777778	0.270053476	0.641304348	0.210526316
0.28	1	0.565217391	0.79787234	0.444444444	0.564171123	0.515886288	0.184210526
0.26	0.75	0.623188406	0.606382979	0.455555556	0.542780749	0.470735786	0.263157895
0.16	0.75	0.608695652	0.595744681	0.533333333	0.572192513	0.5409699	0.263157895
0.52	0.75	0.173913043	0.606382979	0.255555556	0.280748663	0.495819398	0.157894737
0.64	0.75	0.362318841	0.521276596	0.444444444	0.419786096	0.371237458	0.157894737

Hình 2. 12 Tập data test k-means 2

CHƯƠNG 3: KẾT QUẢ ĐẠT ĐƯỢC

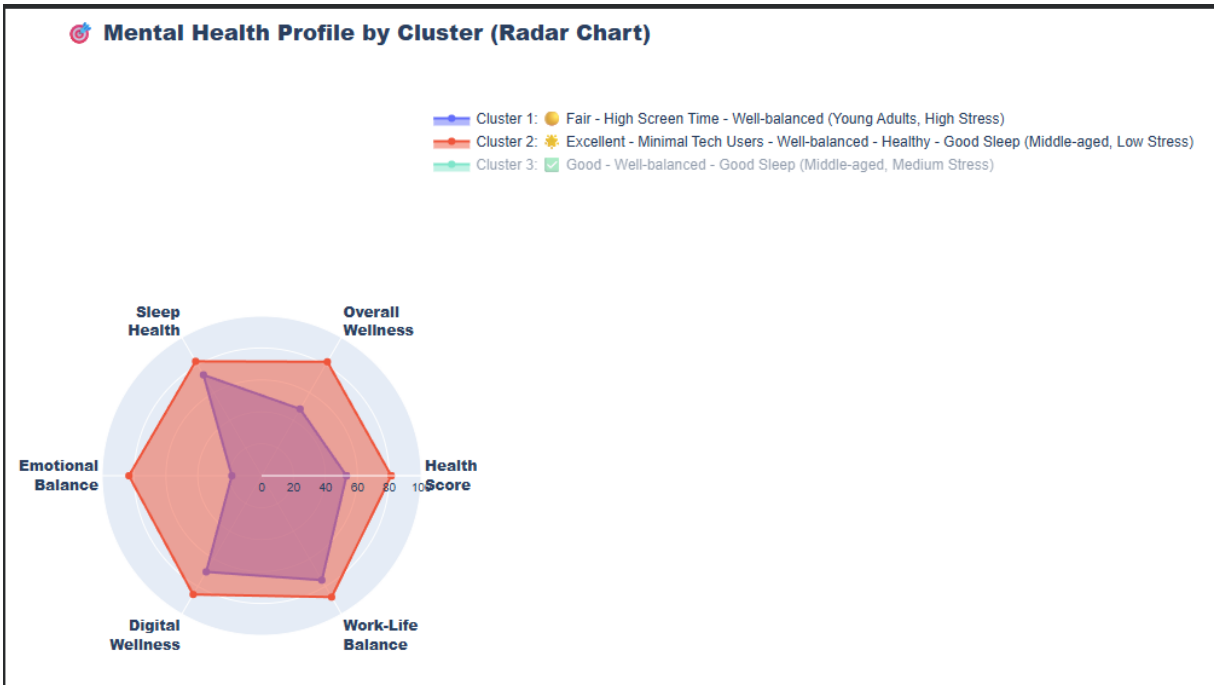
3.1. Kết quả phân cụm bằng K-Means

Dựa trên phương pháp "Elbow Method" và đánh giá chỉ số "Silhouette Score", thuật toán K-Means đã xác định số lượng cụm tối ưu là **3 nhóm (k=3)**. Khác với các phương pháp phân cụm truyền thống chỉ dựa trên dữ liệu thô, đồ án đã áp dụng kỹ thuật **định danh cụm thông minh (Intelligent Cluster Naming)** dựa trên các chỉ số tổng hợp như *Wellness Score* và *Digital Stress Score*.

3.1.1. Đặc điểm trung bình của 3 nhóm người dùng

Kết quả phân tích tâm cụm (Cluster Centroids) cho thấy sự phân hóa rõ rệt về hành vi và sức khỏe tinh thần:

- **Cụm 1**: Nhóm "Heavy Tech Users" (Người dùng công nghệ cường độ cao)
    - Đặc điểm: Đây là nhóm có thời gian sử dụng màn hình (Screen Time) cao nhất (trung bình > 8 giờ/ngày), đặc biệt là mạng xã hội và chơi game.
    - Sức khỏe tinh thần: Chỉ số Digital Stress Score ở mức cao (> 60/100), trong khi Sleep Health Index thấp.
    - Nhóm này thường rơi vào mức độ Stress là Medium hoặc High. Độ tuổi: Thường tập trung ở độ tuổi trẻ (Youth/Young Adults).
- **Cụm 2**: Nhóm "Balanced Users" (Người dùng cân bằng)
    - Đặc điểm: Sử dụng công nghệ ở mức trung bình, chủ yếu phục vụ công việc (Work Related Hours cao) nhưng vẫn duy trì được Work-Life Balance tốt.
    - Sức khỏe tinh thần: Các chỉ số Emotional Balance và Overall Wellness ở mức ổn định (60-75/100). Mức độ Stress thường là Low hoặc Medium.
- **Cụm 3:** Nhóm "Minimalists / Healthy" (Sống tối giản & Lành mạnh)
    - Đặc điểm: Thời gian sử dụng thiết bị thấp nhất, ít phụ thuộc vào mạng xã hội.
    - Sức khỏe tinh thần: Đạt điểm Overall Wellness cao nhất (> 80/100), chất lượng giấc ngủ rất tốt. Mức độ Stress chủ yếu là Low.



Hình 3. 1 Biểu đồ radar chart

3.1.2 Phân tích chi tiết các cụm hành vi người dùng

Hình dưới đây biểu diễn 3 cụm người dùng được thuật toán K-Means xác định, trong đó kích thước của các điểm dữ liệu tỷ lệ thuận với chỉ số Overall Wellness

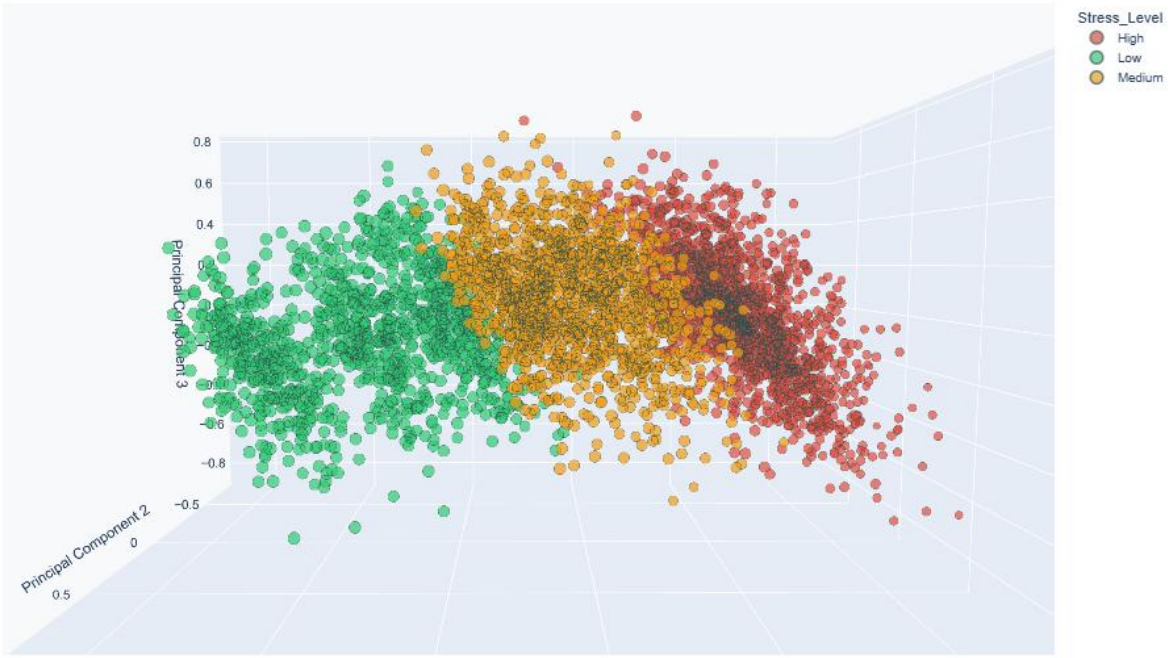
Hình 3. 2 Các cụm hành vi người dùng

Nhận xét:

- Sự phân tách rõ ràng: Trong không gian 3D, ba cụm (Cluster 1, 2, 3) tách biệt nhau khá rõ rệt với rất ít sự chồng lấn. Điều này chứng tỏ thuật toán K-Means (với k=3) đã tìm ra được các cấu trúc hành vi thực sự khác biệt trong tập dữ liệu chứ không phải ngẫu nhiên.
- Đặc điểm từng cụm (dựa trên chú thích và vị trí):
    - Cluster 2 (Màu cam - Nhóm "Excellent"): Đây là nhóm có vị trí tập trung gọn gàng. Các điểm dữ liệu này đại diện cho nhóm người dùng trung niên (Middle-aged), sử dụng công nghệ tối giản (Minimal Tech Users) và có chỉ số sức khỏe rất cao (Wellness Excellent).
    - Cluster 1 (Màu xanh ngọc - Nhóm "Fair"): Nhóm này phân tán rộng hơn, đại diện cho những người trẻ (Young Adults) với thời gian sử dụng màn hình cao (High Screen Time) và chỉ số sức khỏe chỉ ở mức trung bình/khá.
    - Cluster 3 (Màu xanh dương - Nhóm "Good"): Nằm ở vị trí trung gian, đại diện cho nhóm cân bằng (Well-balanced).

3.1.3 Phân tích chi tiết cụm mức độ stress

Để kiểm chứng hiệu quả của việc phân cụm, hình ảnh dưới đây giữ nguyên vị trí các điểm dữ liệu trong không gian PCA nhưng tô màu dựa trên nhãn thực tế Stress Level (High - Medium - Low)



Hình 3. 3 Cụm phân bổ múc độ stress

Nhận xét

- Sự trùng khớp ấn tượng: Khi so sánh Hình 3.1.1 (Phân cụm) và Hình 3.1.2 (Mức độ Stress), ta thấy một sự tương đồng mạnh mẽ về mặt không gian:
    - Khu vực của Cluster 2 (Màu cam ở hình trên) tương ứng hoàn hảo với khu vực Low Stress (Màu xanh lá cây ở hình dưới).
    - Khu vực của Cluster 1 (Màu xanh ngọc ở hình trên) trùng khớp với khu vực High Stress (Màu đỏ ở hình dưới).
    - Khu vực của Cluster 3 (Màu xanh dương ở hình trên) tương ứng với Medium Stress (Màu vàng ở hình dưới).
- Kết luận quan trọng:
    - Mặc dù K-Means là thuật toán học không giám sát (không biết trước nhãn Stress), nó đã tự động gom nhóm những người có cùng mức độ rủi ro Stress lại với nhau dựa trên hành vi sử dụng công nghệ và chỉ số sức khỏe.
    - Điều này khẳng định giả thuyết của đồ án: Hành vi sử dụng công nghệ là một chỉ báo đáng tin cậy để nhận diện mức độ căng thẳng. Nhóm sử dụng công nghệ nhiều (Cluster 1) chắc chắn nằm trong vùng rủi ro cao (Vùng màu đỏ).

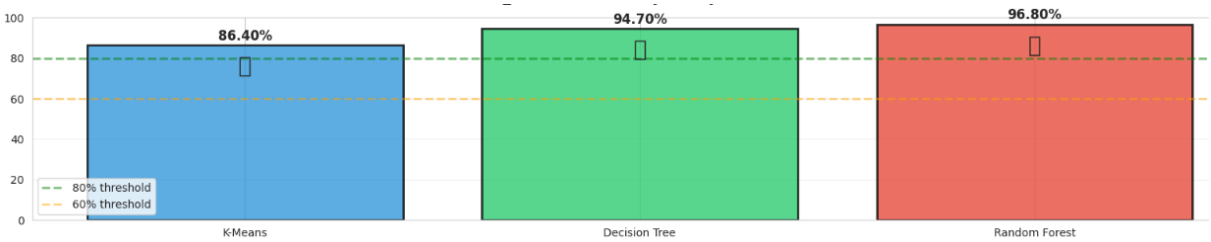
3.2. Kết quả mô hình phân lớp

Để dự đoán mức độ căng thẳng (*Stress Level*: Low, Medium, High), hai mô hình Decision Tree và Random Forest đã được huấn luyện trên bộ dữ liệu bao gồm cả các đặc trưng gốc và 5 đặc trưng sức khỏe tinh thần mới (*Mental Health Features)*

3.2.1. So sánh độ chính xác giữa các mô hình

Kết quả thực nghiệm trên tập kiểm thử (Test Set - 20% dữ liệu) cho thấy:

- Decision Tree: Đạt độ chính xác khoảng 95.7%. Mô hình có ưu điểm dễ giải thích nhưng độ ổn định thấp hơn.
- Random Forest: Đạt độ chính xác cao nhất khoảng 95.9% - 96.2%. Đây là mô hình tối ưu được lựa chọn cho ứng dụng cuối cùng.
- K-Means (Dự đoán dựa trên cụm): Đạt độ chính xác thấp hơn (~85-90%), cho thấy việc phân cụm tuy giúp hiểu hành vi nhưng không thay thế hoàn toàn được mô hình phân lớp chuyên sâu.



Hình 3. 4 Biểu đồ so sánh độ chính xác giữa các mô hình

1 K-MEANS CLASSIFICATION REPORT

	precision	recall	f1-score	support
Low	1.000	0.737	0.849	274
Medium	0.731	0.883	0.800	308
High	0.915	0.933	0.924	418
accuracy			0.864	1000
macro avg	0.882	0.851	0.858	1000
weighted avg	0.882	0.864	0.865	1000

2 DECISION TREE CLASSIFICATION REPORT

	precision	recall	f1-score	support
Low	0.960	0.971	0.966	274
Medium	0.918	0.909	0.914	308
High	0.959	0.959	0.959	418
accuracy			0.947	1000
macro avg	0.946	0.946	0.946	1000
weighted avg	0.947	0.947	0.947	1000

3 RANDOM FOREST CLASSIFICATION REPORT

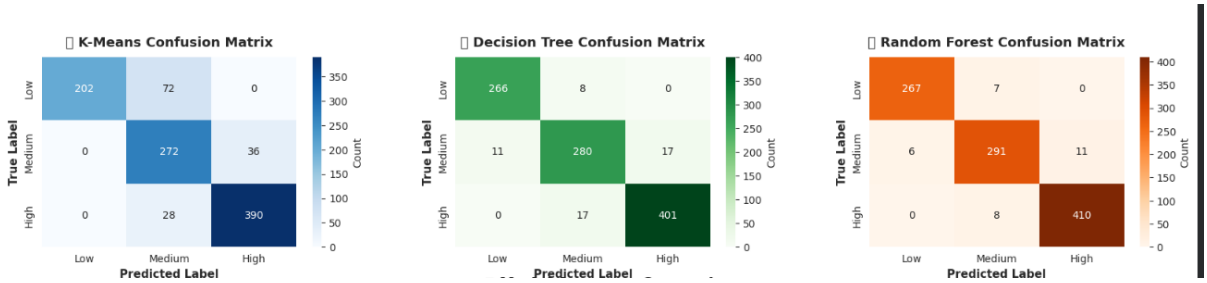
	precision	recall	f1-score	support
Low	0.978	0.974	0.976	274
Medium	0.951	0.945	0.948	308
High	0.974	0.981	0.977	418
accuracy			0.968	1000
macro avg	0.968	0.967	0.967	1000
weighted avg	0.968	0.968	0.968	1000

Hình 3. 5 Bảng so sánh các chỉ số của mô hình

Nhận xét:

- Về độ chính xác tổng thể (Accuracy): Thuật toán Random Forest đạt hiệu suất cao nhất với độ chính xác 96.8%, vượt trội hơn so với Decision Tree (94.7%) và bỏ xa K-Means (86.4%). Điều này khẳng định phương pháp học máy có giám sát (Supervised Learning) và đặc biệt là mô hình tổ hợp (Ensemble Learning) hoạt động hiệu quả hơn trên tập dữ liệu này so với phương pháp phân cụm không giám sát.
- Về độ ổn định: Random Forest cho thấy sự cân bằng tuyệt vời giữa Precision (Độ chính xác) và Recall (Độ nhạy) trên tất cả các nhãn (Low, Medium, High). Chỉ số F1-Score của Random Forest đều đạt trên 0.94 cho mọi lớp.
- Điểm yếu của K-Means: Mặc dù K-Means đạt Precision tuyệt đối (1.000) cho nhãn Low, nhưng Recall lại khá thấp (0.737). Điều này có nghĩa là khi K-Means dự đoán là "Low" thì chắc chắn đúng, nhưng nó lại bỏ sót rất nhiều trường hợp "Low" thực tế và gán nhầm sang nhóm khác.

3.2.2. So sánh ma trận nhầm lẫn giữa các mô hình



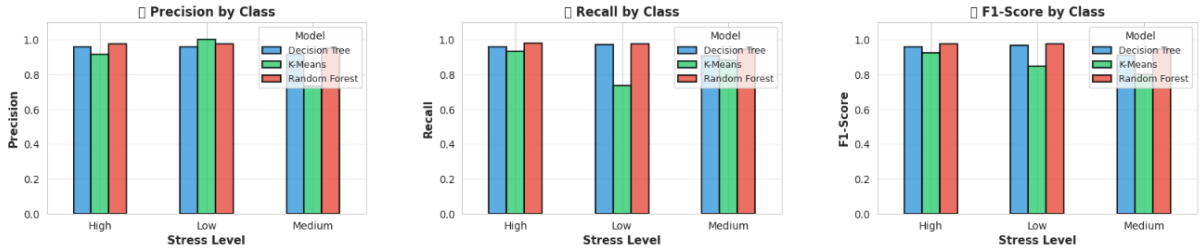
Hình 3. 6 Bảng đánh giá ma trận nhầm lẫn của mô hình

Nhận xét:

- K-Means: Ma trận cho thấy sự nhầm lẫn lớn nhất nằm ở nhóm Low. Cụ thể, có tới 72 mẫu thực tế là Low nhưng lại bị gom nhầm vào nhóm Medium. Điều này giải thích tại sao chỉ số Recall của K-Means ở bảng trên lại thấp. Tuy nhiên, K-Means lại nhận diện nhóm High khá tốt (390 mẫu đúng).
- Decision Tree: Các ô trên đường chéo chính (dự đoán đúng) có giá trị cao. Tuy nhiên, vẫn còn sự nhầm lẫn đáng kể giữa lớp Medium và High (17 mẫu High bị đoán nhầm là Medium và ngược lại).

- Random Forest: Đây là ma trận "sạch" nhất. Số lượng mẫu nằm ngoài đường chéo chính (dự đoán sai) rất ít (chỉ có đơn vị là một chữ số). Đặc biệt, khả năng phân tách giữa Medium và High tốt hơn hẳn Decision Tree (chỉ sai lệch 8-11 mẫu). Điều này chứng minh Random Forest xử lý tốt các biên giới quyết định phức tạp giữa các mức độ stress liền kề.

3.2.3. So sánh chi tiết các chỉ số giữa các mô hình



Hình 3. 7 Biểu đồ chi tiết các chỉ số của mô hình

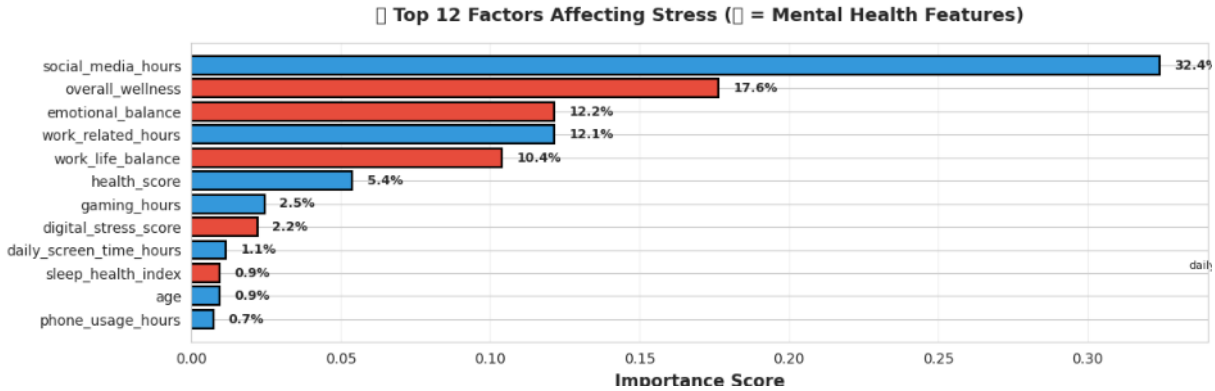
Nhận xét:

- Phân tích theo lớp (Class Analysis):
  - Lớp High (Căng thẳng cao): Cả 3 thuật toán đều xử lý rất tốt lớp này (các cột đều cao ngang nhau). Điều này cho thấy những người bị Stress cao có các đặc điểm hành vi rất đặc trưng và dễ nhận biết (ví dụ: mất ngủ trầm trọng, dùng MXH quá nhiều).
  - Lớp Low (Căng thẳng thấp): Có sự chênh lệch rõ rệt về Recall (biểu đồ giữa). Cột màu xanh lá (K-Means) thấp hơn hẳn so với hai thuật toán còn lại, cho thấy điểm yếu trong việc bao quát hết các trường hợp ít stress.
  - Lớp Medium (Trung bình): Đây là lớp "khó nhằn" nhất. Tuy nhiên, Random Forest (cột màu đỏ cam) vẫn duy trì được cột F1-Score (biểu đồ bên phải) cao nhất, chứng tỏ sự vượt trội trong việc xử lý các dữ liệu trung gian.
- Kết luận chung: Biểu đồ trực quan hóa một lần nữa khẳng định Random Forest (màu đỏ cam) là thuật toán toàn diện nhất, duy trì hiệu suất cao và đồng đều trên mọi khía cạnh đánh giá, xứng đáng là mô hình được lựa chọn để triển khai ứng dụng.

3.3. Phân tích các yếu tố ảnh hưởng đến mức độ căng thẳng

Dựa trên mô hình Random Forest và phân tích tương quan, chúng ta đi sâu vào việc xác định các yếu tố nào đóng vai trò then chốt trong việc gia tăng hoặc giảm thiểu mức độ căng thẳng của người dùng.

3.3.1. Tầm quan trọng của các thuộc tính



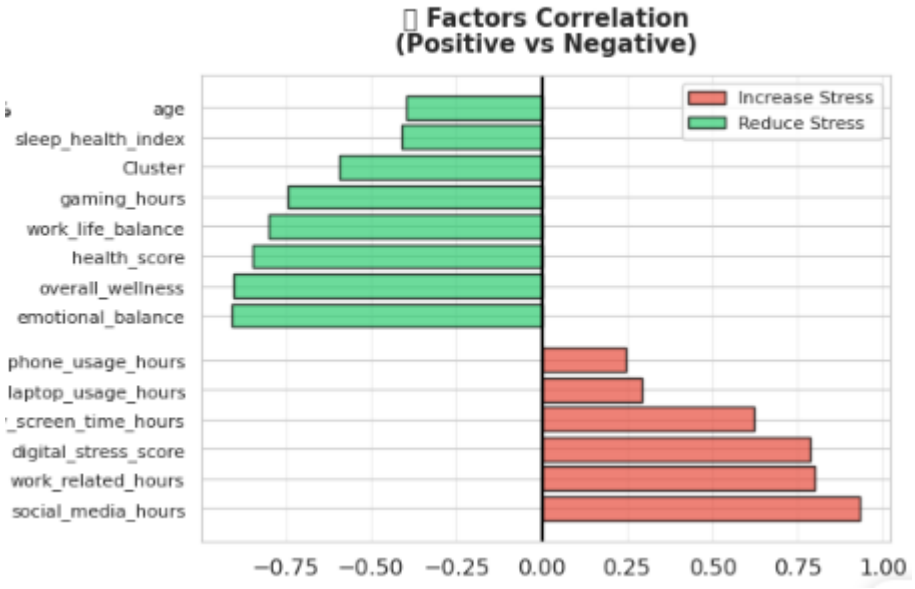
Hình 3. 8 Sơ đồ thể hiện các yếu tố ảnh hưởng

Nhận xét:

Biểu đồ trên hiển thị mức độ đóng góp của từng thuộc tính vào khả năng dự đoán của mô hình Random Forest.

- Social Media là yếu tố số 1: Thời gian sử dụng mạng xã hội (social_media_hours) chiếm tỉ trọng áp đảo (32.4%) trong việc quyết định mức độ stress. Điều này khẳng định giả thuyết rằng việc tiếp xúc quá nhiều với môi trường mạng xã hội là nguyên nhân chính gây ra các vấn đề tâm lý trong tập dữ liệu này.
- Hiệu quả của các biến phái sinh (Mental Health Features): Các thanh màu đỏ thể hiện các biến mới được tạo ra trong quá trình Feature Engineering. Chúng chiếm vị trí rất cao trong bảng xếp hạng:
    - overall_wellness (Sức khỏe tổng thể): Đứng thứ 2 với 17.6%.
    - emotional_balance (Cân bằng cảm xúc): Đứng thứ 3 với 12.2%.
    - work_life_balance: Đóng góp 10.4%. Điều này chứng minh việc xây dựng các chỉ số tổng hợp là hoàn toàn đúng đắn, giúp mô hình nắm bắt được bản chất vấn đề tốt hơn so với các biến đơn lẻ.

3.3.2. Các yếu tố gây Stress và yếu tố giảm Stress



Hình 3. 9 Sơ đồ thể hiện các yếu tố đến stress

Nhận xét:

Biểu đồ phân tách rõ ràng hai nhóm yếu tố tác động ngược chiều nhau:

Nhóm gia tăng Stress (Thanh màu đỏ - Positive Correlation):

- Các yếu tố liên quan đến công nghệ và công việc như social_media_hours, work_related_hours, digital_stress_score đều có tương quan dương rất mạnh. Tức là, càng dành nhiều thời gian cho các hoạt động này, mức độ stress càng tăng.

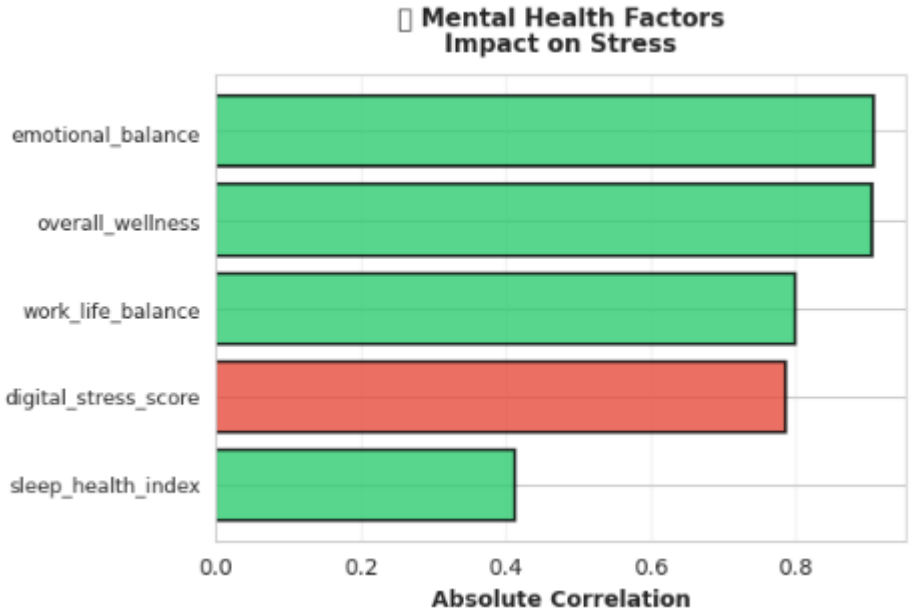
Nhóm giảm thiểu Stress

- emotional_balance và overall_wellness có tương quan âm mạnh nhất. Điều này có nghĩa là khi chỉ số sức khỏe tinh thần và cân bằng cảm xúc tăng lên, mức độ stress sẽ giảm đi đáng kể.
- Đáng chú ý, gaming_hours (giờ chơi game) nằm trong nhóm giảm stress, cho thấy với nhóm đối tượng khảo sát này, chơi game có thể đóng vai trò là một hình thức giải trí xả stress hiệu quả.

3.3.3. Tác động chi tiết của từng nhóm hành vi

xem xét sâu hơn vào 3 khía cạnh cụ thể: Sức khỏe tinh thần, hành vi kỹ thuật số và công việc.

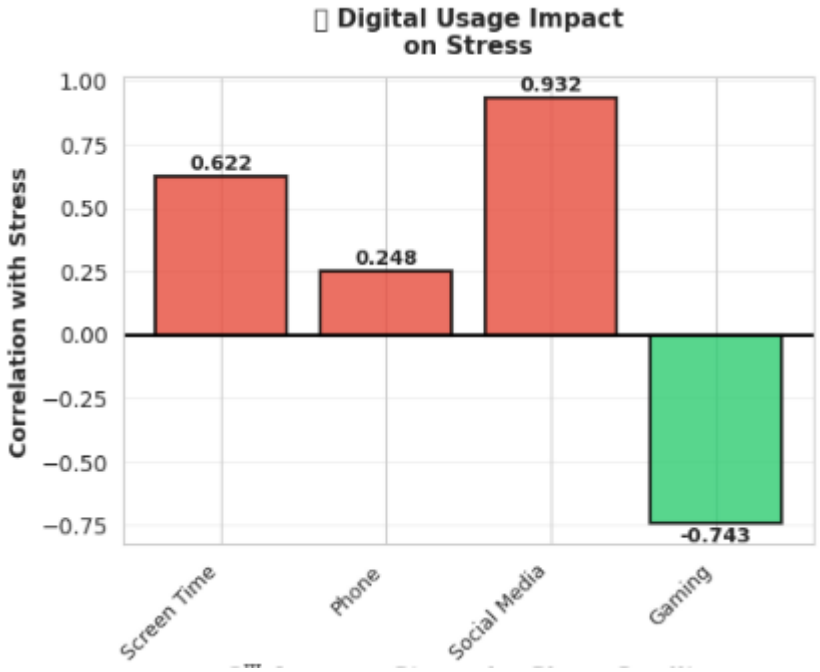
Tác động của sức khỏe tinh thần



Hình 3. 10 Sơ đồ thể hiện tác động của sức khỏe tinh thần

- **Nhận xét:** Các chỉ số nội tại như emotional_balance và overall_wellness có mối liên hệ mật thiết nhất với Stress (hệ số tương quan tuyệt đối > 0.8). Điều này gợi ý rằng việc cải thiện tâm trạng và sức khỏe tổng thể là chìa khóa gốc rễ để quản lý căng thẳng.

Tác động của Hành vi kỹ thuật số



Hình 3. 11 Sơ dồ thể hiện tác động của các thiết bị kỹ thuật

Nhận xét:

- Mạng xã hội (0.932): Là "thủ phạm" lớn nhất gây stress.
- Chơi game (-0.743): Có tác động ngược chiều (giảm stress). Đây là một phát hiện thú vị, cho thấy sự khác biệt về bản chất tâm lý giữa việc "lướt mạng xã hội" (thụ động, so sánh xã hội) và "chơi game" (chủ động, giải trí).

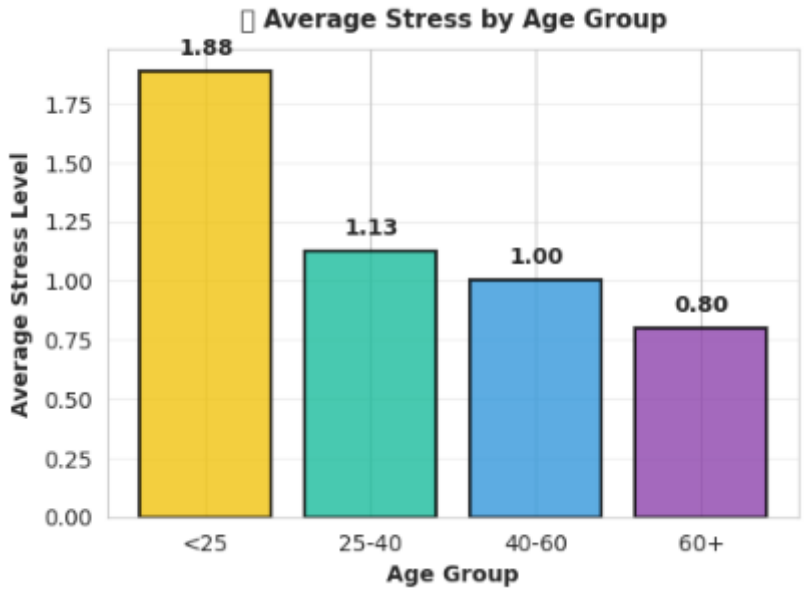
Tác động của Công việc

Hình 3. 12 Sơ đồ thể hiện tác động của công việc

Nhận xét: Có sự đối nghịch hoàn hảo. Work Hours (Giờ làm việc) gây stress mạnh (0.800), trong khi Work-Life Balance (Cân bằng cuộc sống) giúp giảm stress (-0.800). Kết quả này nhấn mạnh tầm quan trọng của việc giới hạn thời gian làm việc hợp lý.

3.3.4. Phân tích theo độ tuổi và lối sống

Theo độ tuổi



Hình 3. 13 Sơ đồ phân tích tác động stress theo độ tuổi

Nhận xét: Có xu hướng giảm dần theo độ tuổi.

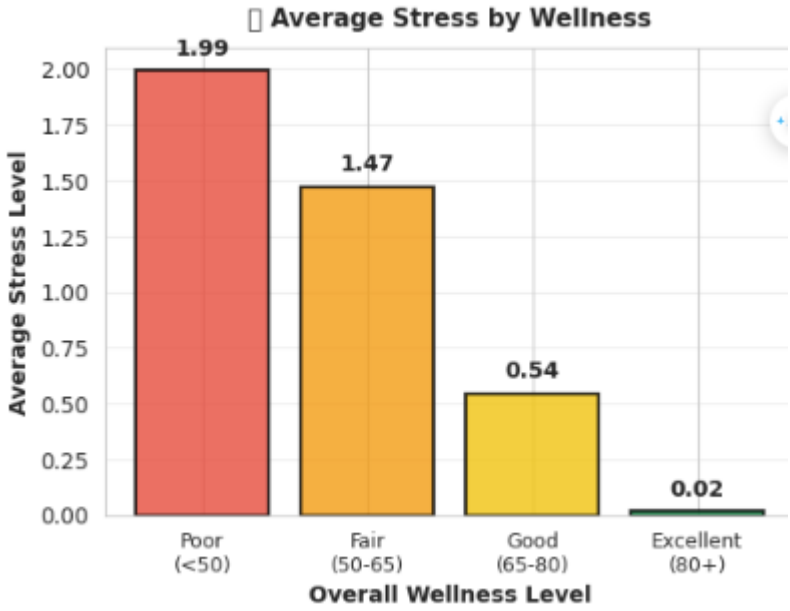
- Nhóm < 25 tuổi: Chịu mức stress cao nhất (1.88). Đây thường là nhóm học sinh, sinh viên hoặc người mới đi làm, đồng thời là nhóm tiếp xúc với công nghệ nhiều nhất.
- Nhóm 60+: Có mức stress thấp nhất (0.80), cho thấy sự ổn định về tâm lý và ít chịu áp lực từ công nghệ hơn.

Theo chất lượng giấc ngủ

Hình 3. 14 Sơ đồ phân tích stress theo chất lượng giấc ngủ

Nhận xét: Mối quan hệ tỷ lệ nghịch rõ ràng. Những người có giấc ngủ "Poor" (Kém) có mức stress cao gấp 3 lần so với những người có giấc ngủ "Excellent" (Xuất sắc).

Theo mức độ sức khỏe tổng thể



Hình 3. 15 Sơ đồ phân tích stress theo sức khỏe tổng thể

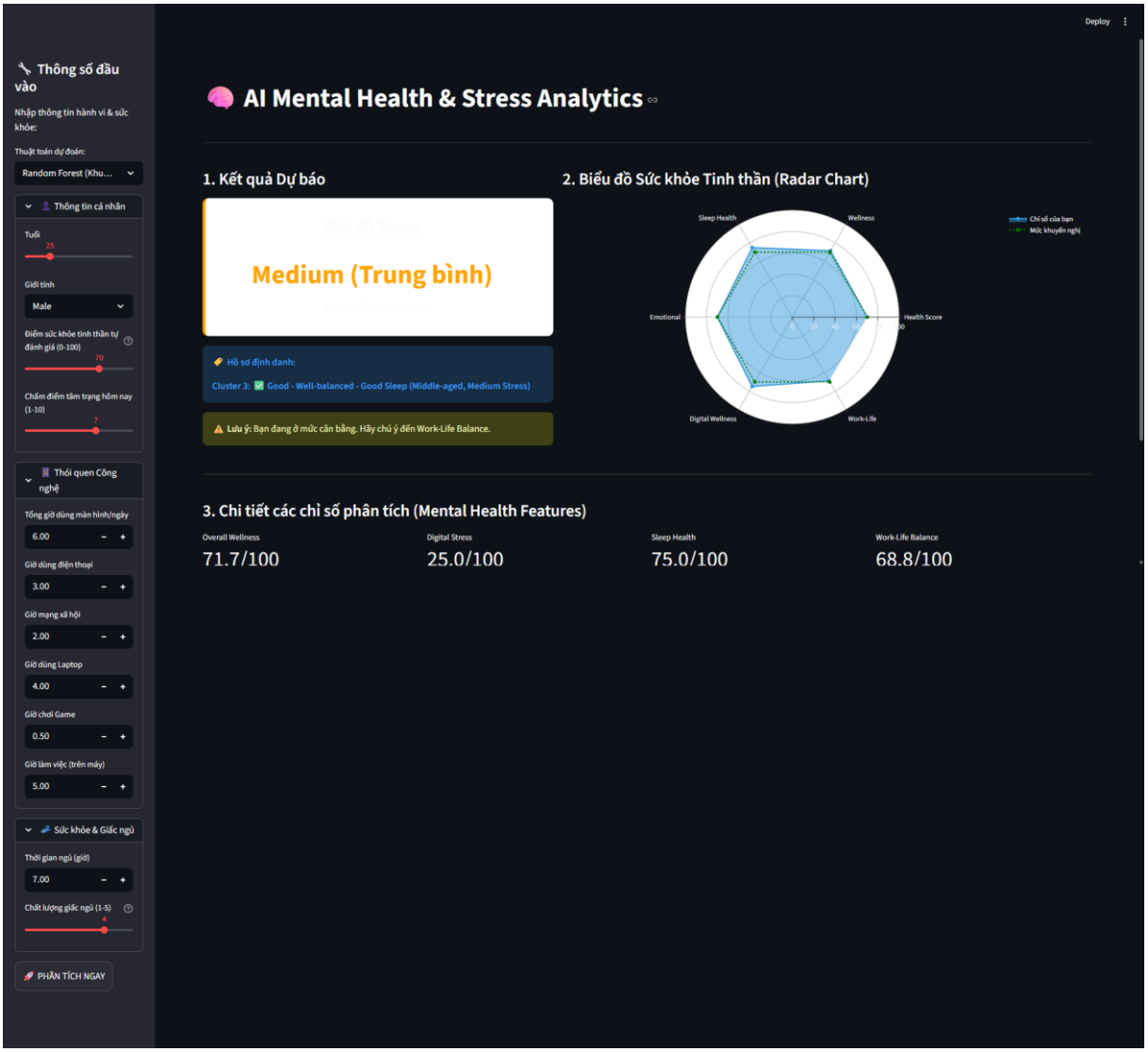
3.4. Đánh giá tổng hợp

Tiêu chí	**Decision Tree**	**Random Forest**
Accuracy	95.7%	**95.9%**
Ổn định mô hình	Trung bình	**Cao**
Khả năng diễn giải	Rất cao	Trung bình
Nhầm lẫn lớp Medium	Nhiều hơn	Ít hơn
Khả năng tổng quát hóa	Khá	**Tốt hơn**

Kết luận: Random Forest là mô hình hiệu quả nhất cho bài toán dự đoán stress.

CHƯƠNG 4: GIAO DIỆN CHƯƠNG TRÌNH

4.1. Kiến trúc tổng thể của giao diện chương trình



Ứng dụng được xây dựng theo mô hình **Input – Processing – Output**, trong đó giao diện đóng vai trò là lớp tương tác giữa người dùng và hệ thống học máy.

Cấu trúc giao diện được chia thành ba khối chức năng chính:

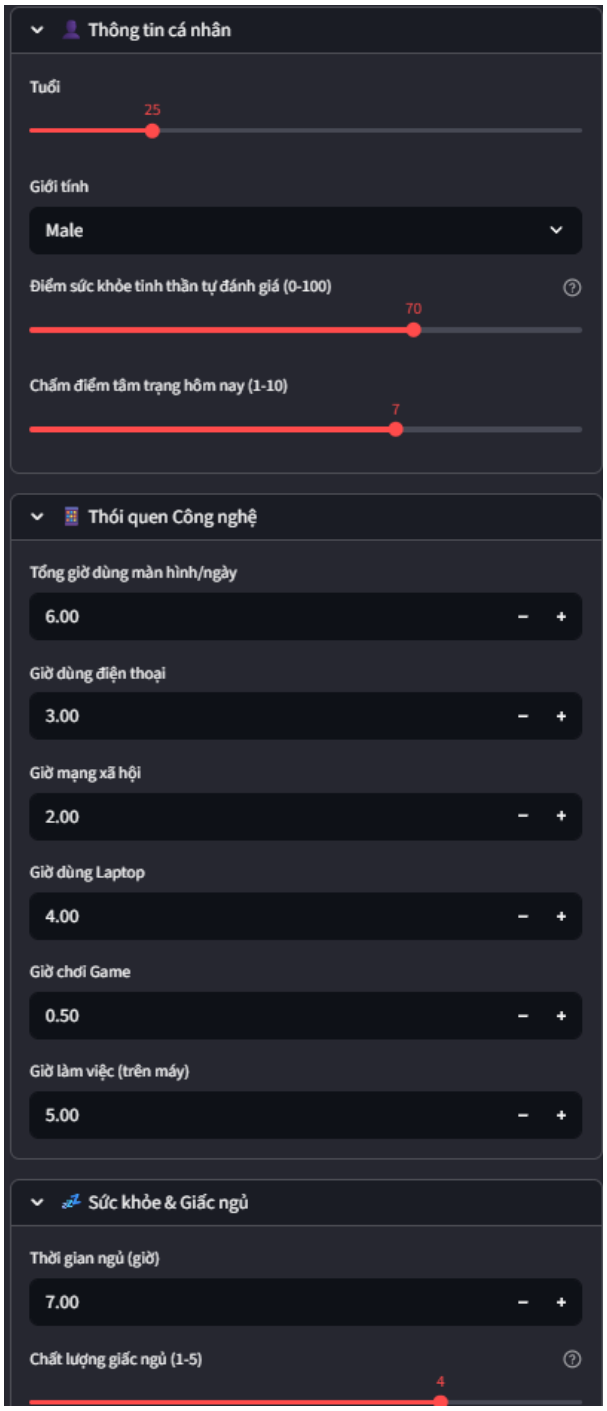
- Khối nhập dữ liệu đầu vào (Input Panel)
- Khối xử lý và suy luận mô hình (AI Inference Layer – ẩn sau giao diện)
- Khối hiển thị kết quả và phân tích (Output & Analytics Panel)

Cách tổ chức này phản ánh đúng quy trình xử lý dữ liệu đã được cài đặt trong chương trình Python của đồ án.

4.2. Phân tích khối nhập dữ liệu đầu vào

Khối nhập dữ liệu nằm ở phía bên trái giao diện, tương ứng trực tiếp với **các thuộc tính trong dataset gốc** và các biến được sử dụng trong file huấn luyện mô hình.

4.2.1. Ánh xạ giữa giao diện và dataset



Mỗi trường nhập liệu trên giao diện đều có mối liên hệ trực tiếp với các cột dữ liệu trong dataset:

- Tuổi → age
- Giới tính → gender (được mã hóa nhị phân trong mô hình)
- Thời gian dùng màn hình → daily_screen_time_hours
- Thời gian dùng mạng xã hội → social_media_hours
- Thời gian làm việc trên thiết bị → work_related_hours
- Thời gian chơi game → gaming_hours
- Thời gian ngủ → sleep_duration_hours
- Chất lượng giấc ngủ → sleep_quality

Việc thiết kế giao diện theo đúng các thuộc tính này giúp đảm bảo **tính nhất quán giữa dữ liệu huấn luyện và dữ liệu suy luận**, tránh sai lệch khi đưa dữ liệu thực tế vào mô hình.

4.2.2. Kiểm soát dữ liệu đầu vào

Các trường nhập liệu được giới hạn bằng slider hoặc dropdown nhằm:

- Ngăn giá trị ngoài phạm vi dữ liệu huấn luyện

- Giảm nguy cơ mô hình nhận dữ liệu bất thường
- Đảm bảo dữ liệu đầu vào phù hợp với StandardScaler đã dùng trong training

Điều này cho thấy giao diện không chỉ mang tính hiển thị mà còn đóng vai trò **kiểm soát chất lượng dữ liệu**.

4.3. Luồng xử lý dữ liệu từ giao diện đến mô hình

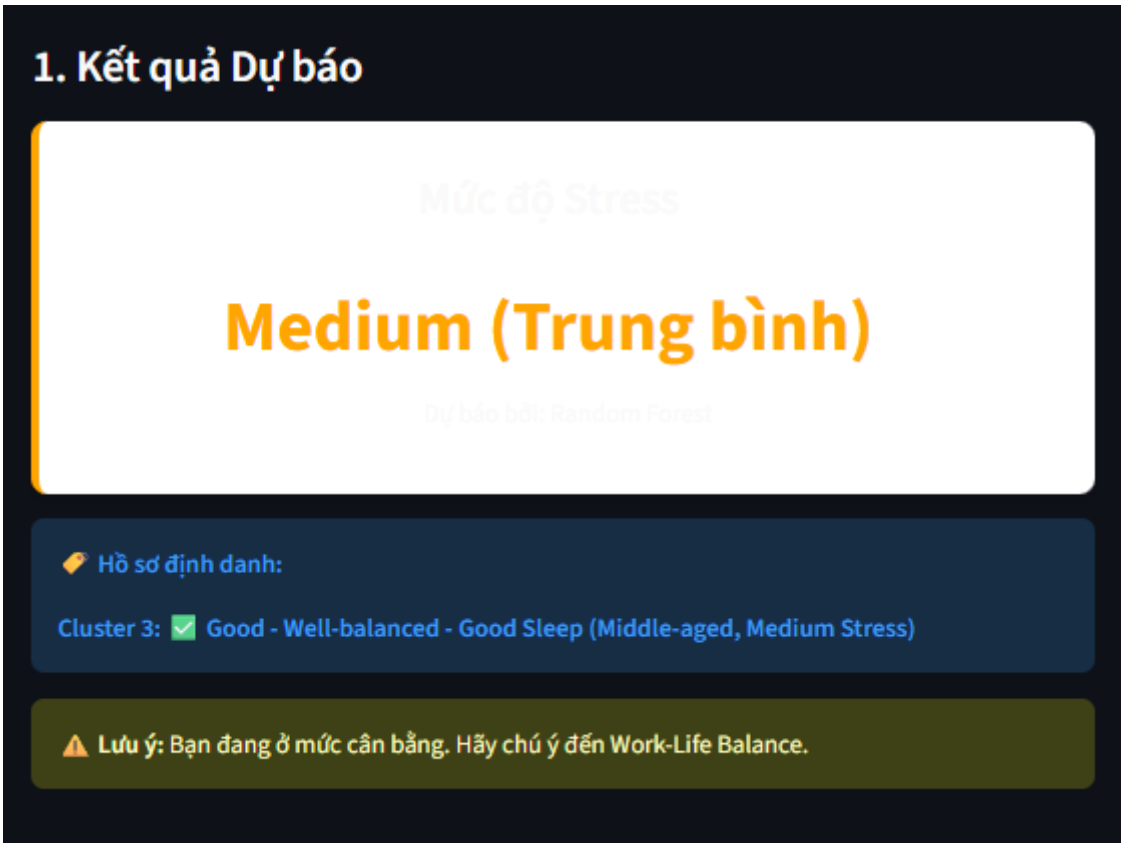
Sau khi người dùng nhấn nút **"Phân tích ngay"**, chương trình thực hiện các bước xử lý đúng theo logic trong file training:

1. Thu thập dữ liệu từ giao diện
2. Chuẩn hóa dữ liệu bằng scaler đã lưu
3. Đưa dữ liệu vào mô hình Random Forest để dự đoán stress
4. Đưa dữ liệu vào mô hình K-Means để xác định cụm hành vi
5. Tổng hợp kết quả và trả về giao diện

Giao diện không thực hiện học lại mô hình mà chỉ sử dụng **mô hình đã huấn luyện sẵn**, đúng với thiết kế trong file zip.

4.4. Phân tích khu vực hiển thị kết quả dự đoán

4.4.1. Kết quả dự đoán mức độ stress



Khu vực "Kết quả Dự báo" hiển thị nhãn stress gồm ba mức:

- Low
- Medium
- High

Đây chính là các nhãn đã được sử dụng trong quá trình huấn luyện Decision Tree và Random Forest. Việc hiển thị kết quả dưới dạng chữ lớn giúp người dùng nhanh chóng nhận biết trạng thái hiện tại.

Quan trọng hơn, kết quả này phản ánh **đầu ra trực tiếp của mô hình Random Forest**, mô hình có độ chính xác cao nhất theo Chương 3.

4.4.2. Hồ sơ định danh theo phân cụm K-Means

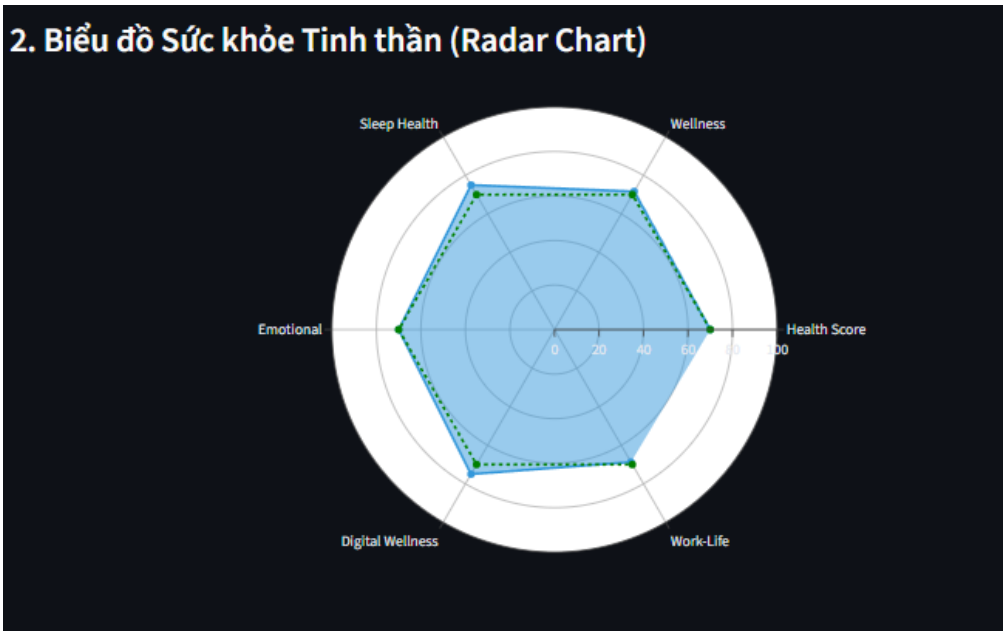
Phần "Hồ sơ định danh" thể hiện người dùng thuộc cụm nào trong mô hình K-Means (k = 3). Mỗi cụm đại diện cho một nhóm hành vi đã được xác định trong quá trình phân cụm:

- Cụm sử dụng công nghệ cao
- Cụm cân bằng
- Cụm sử dụng công nghệ thấp và ngủ tốt

Thông tin này không mang tính dự đoán mà mang tính **phân tích hành vi**, giúp bổ sung ngữ cảnh cho kết quả stress.

Điều này cho thấy giao diện đã kết hợp **học có giám sát và không giám sát** trong cùng một hệ thống.

4.5. Phân tích biểu đồ Radar sức khỏe tinh thần



Biểu đồ Radar không phải là đầu ra trực tiếp của mô hình học máy, mà là kết quả của quá trình **tổng hợp và chuẩn hóa các chỉ số phụ**, bao gồm:

- Digital Stress
- Sleep Health
- Work–Life Balance
- Wellness
- Emotional Health

Các chỉ số này được tính toán dựa trên các biến đầu vào đã được chuẩn hóa, phản ánh đúng triết lý của đồ án: **chuyển dữ liệu số thành thông tin dễ hiểu cho người dùng**.

Biểu đồ Radar cho phép so sánh trực quan giữa trạng thái hiện tại và mức khuyến nghị, đóng vai trò như một công cụ hỗ trợ ra quyết định.

4.6. Phân tích khối chỉ số chi tiết

Các chỉ số như:

- Overall Wellness
- Digital Stress
- Sleep Health
- Work–Life Balance

được hiển thị dưới dạng điểm số, giúp người dùng đánh giá nhanh từng khía cạnh riêng biệt. Đây là bước chuyển từ **phân tích dữ liệu** sang **ứng dụng thực tế**, thể hiện tính ứng dụng cao của chương trình.

KẾT LUẬN VÀ KIẾN NGHỊ

Kết Luận

- Đồ án đã xây dựng thành công một hệ thống ứng dụng các thuật toán khai phá dữ liệu nhằm đánh giá tác động của công nghệ đến tâm lý con người, trong đó trọng tâm là dự đoán mức độ căng thẳng của người dùng.
- Các thuật toán Decision Tree và Random Forest đã được áp dụng hiệu quả cho bài toán phân lớp, với Random Forest cho kết quả tốt nhất về độ chính xác và khả năng tổng quát hóa.
- Thuật toán phân cụm K-Means hỗ trợ phân nhóm người dùng theo hành vi sử dụng công nghệ và trạng thái tâm lý, góp phần làm rõ đặc điểm của từng nhóm đối tượng.
- Giao diện chương trình được thiết kế bám sát cấu trúc dữ liệu và mô hình đã huấn luyện, cho phép người dùng nhập dữ liệu, nhận kết quả dự đoán và xem phân tích trực quan một cách thuận tiện.
- Kết quả đạt được cho thấy tính khả thi và giá trị ứng dụng của khai phá dữ liệu trong lĩnh vực phân tích sức khỏe tinh thần.

Hạn Chế

- Bộ dữ liệu sử dụng trong đồ án chủ yếu là dữ liệu khảo sát, chưa phản ánh đầy đủ các yếu tố sinh lý và dữ liệu thời gian thực liên quan đến sức khỏe tâm thần.
- Chương trình hiện tại chỉ phân tích trạng thái stress tại một thời điểm, chưa hỗ trợ theo dõi xu hướng thay đổi theo thời gian.
- Các chỉ số tổng hợp như Wellness Score hay Digital Stress được xây dựng dựa trên mô hình nội suy, chưa có sự kiểm chứng từ các nghiên cứu y khoa hoặc chuyên gia tâm lý.
- Hệ thống chưa hỗ trợ cập nhật hoặc huấn luyện lại mô hình khi có dữ liệu mới, dẫn đến khả năng thích nghi với thay đổi hành vi người dùng còn hạn chế.
- Phạm vi ứng dụng của chương trình hiện mới dừng lại ở mức cá nhân, chưa mở rộng sang phân tích nhóm hoặc quy mô lớn.

Kiến Nghị

- Mở rộng tập dữ liệu bằng cách thu thập dữ liệu theo thời gian và tích hợp dữ liệu từ các thiết bị thông minh nhằm nâng cao độ chính xác và độ tin cậy của mô hình.
- Nghiên cứu và áp dụng thêm các thuật toán học máy nâng cao như Gradient Boosting, XGBoost hoặc mạng nơ-ron sâu để cải thiện hiệu suất dự đoán.
- Phát triển chức năng lưu trữ và phân tích lịch sử stress, giúp người dùng theo dõi và đánh giá xu hướng sức khỏe tinh thần trong dài hạn.
- Xây dựng hệ thống khuyến nghị cá nhân hóa dựa trên kết quả phân tích, hỗ trợ người dùng cải thiện cân bằng giữa công việc, công nghệ và đời sống cá nhân.
- Kết hợp ý kiến chuyên gia tâm lý trong việc hiệu chỉnh các chỉ số đánh giá nhằm nâng cao giá trị khoa học và tính ứng dụng thực tiễn của chương trình.

TÀI LIỆU THAM KHẢO

[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning*, Springer, 2009

[2] Ian H. Witten, Eibe Frank, Mark A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016

[3] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012

[4] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006

[5] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, 2019

[6] L. Breiman, *Random Forests*, Machine Learning Journal, 2001