ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

📖 📖 📖

LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

TEAM PLAN AND WORK DISTRIBUTION

Nhóm 02: Phân tích và dự đoán nguy cơ đột quỵ

Thành phố Hồ Chí Minh, ngày 27 tháng 12 năm 2025

MỤC LỤC

I. Thông tin nhóm

- **Mã nhóm:** 02
- **Tên nhóm:** TRINITY US
- **Danh Sách Thành Viên**

STT	Họ và Tên	MSSV	Email	Vai trò chính
1	Bùi Đoàn Thúy Vy	22120448	22120448@student.hcmus.edu.vn	Q1, Q2, EDA
2	Phạm Trần Trung Hậu	22120100	22120100@student.hcmus.edu.vn	Q3, Q4, Preprocessing
3	Nguyễn Việt Hoàng	22120113	22120113@student.hcmus.edu.vn	Q5, Q6, ML

II. Phân Công Công Việc Chi Tiết

1. Bùi Đoàn Thúy Vy

Công việc chính:

Q1 - Young vs Old Stroke Patterns (04_Q1_Young_vs_Old_Stroke_Patterns.ipynb)

- Phân tích so sánh patterns giữa người trẻ (< 40 tuổi) và người già (> 60 tuổi)
- Xác định yếu tố lối sống vs bệnh nền cho từng nhóm
- Tạo visualizations so sánh chi tiết
- Viết Motivation & Benefits
- Rút ra insights và khuyến nghị

Q2 - Urban Health vs Lifestyle Paradox (05_Q2_Urban_Health_vs_Lifestyle_Paradox.ipynb)

- Phân tích paradox: Sức khỏe tốt + Lối sống xấu vs Sức khỏe xấu + Lối sống tốt
- Stratified analysis theo nhóm đô thị
- Tạo biểu đồ phức tạp cho multi-dimensional comparisons
- Viết Motivation & Benefits

- Đưa ra khuyến nghị cho chính sách y tế

Công việc hỗ trợ:

- **EDA (02_EDA.ipynb):** Hỗ trợ phân tích khám phá dữ liệu
- **Visualization Framework:** Thiết kế style và color scheme nhất quán
- **Code Review:** Review code của các thành viên khác
- **Documentation:** Viết comments và markdown giải thích

Đóng góp chính:

- Phát hiện "Health-Lifestyle Paradox"
- Thiết kế visualization framework cho toàn bộ dự án
- Viết documentation chi tiết cho Q1 và Q2

2. Phạm Trần Trung Hậu

Công việc chính:

Q3 - Smoking × Age Interaction (06_Q3_Smoking_Age_Interaction.ipynb)

- Phân tích tương tác giữa hút thuốc và tuổi tác
- Stratified analysis theo nhóm tuổi
- Statistical testing cho interaction effects
- Visualization cho 3-way interactions (smoking × age × stroke)
- Viết Motivation & Benefits
- Rút ra insights cho chiến dịch cai thuốc

Q4 - Glucose × Gender Interaction (07_Q4_Glucose_Gender_Interaction.ipynb)

- Phân tích tương tác giữa glucose và giới tính
- So sánh sensitivity của nam và nữ với glucose cao
- Statistical testing (ANOVA, Chi-square)
- Visualization comparisons
- Viết Motivation & Benefits
- Đề xuất tiêu chuẩn khác nhau cho nam/nữ

Công việc hỗ trợ:

- **Data Preprocessing (03_Data_Preprocessing.ipynb):** Hỗ trợ xử lý dữ liệu
- **Feature Engineering:** Tạo các features mới cho phân tích
- **Statistical Testing:** Implement các test thống kê
- **Code Review:** Review và test code

Đóng góp chính:

- Implement interaction analysis framework
- Phát hiện gender differences trong glucose sensitivity
- Build data preprocessing pipeline

3. Nguyễn Việt Hoàng

Công việc chính:

Q5 - Urban Rural Heart Disease (08_Q5_Urban_Rural_Heart_Disease.ipynb)

- Phân tích Urban Penalty trong nhóm có bệnh tim
- Stratified analysis theo giới tính
- Phát hiện và sửa lỗi (urban penalty ảnh hưởng nam > nữ)
- Viết Statistical Explanation cho p > 0.05
- Viết Limitations & Future Directions chi tiết
- Áp dụng Precautionary Principle

Q6 - ML Behavioral Persona Prediction (09_Q6_ML_Behavioral_Persona_Prediction.ipynb)

- Tạo 6 behavioral personas bằng K-Means clustering
- Implement 3 ML models: Logistic Regression, Random Forest, XGBoost
- Xử lý class imbalance với SMOTE
- Feature Importance analysis
- Model comparison và evaluation
- Viết Motivation & Benefits

- Viết Limitations & Future Directions (tiếng Việt)

Công việc hỗ trợ:

- **Machine Learning:** Setup ML pipeline, model training
- **Project Coordination:** Quản lý timeline, tích hợp các phần
- **Quality Assurance:** Kiểm tra consistency, sửa lỗi
- **Final Review:** Review toàn bộ dự án trước khi nộp

Đóng góp chính:

- Phát hiện "Urban Penalty" pattern (nam +6.7%, nữ +0.2%)
- Implement Behavioral Persona ML model
- Viết Statistical vs Clinical Significance explanation
- Điều phối timeline và deliverables

4. Công việc chung

Tất cả thành viên cùng tham gia:

1. Data Collection (01_Data_Collection.ipynb)

- Thu thập và tìm hiểu dataset
- Download từ Kaggle
- Khám phá cấu trúc dữ liệu ban đầu

2. EDA (02_EDA.ipynb)

- Exploratory Data Analysis
- Phân tích phân phối các biến
- Phát hiện patterns ban đầu
- Tạo visualizations tổng quan

3. Data Preprocessing (03_Data_Preprocessing.ipynb)

- Xử lý missing values
- Encoding categorical variables
- Feature engineering

- Tạo healthcare_cleaned.csv

4. Conclusion Summary (10_Conclusion_Summary.ipynb)

- Tổng hợp findings từ 6 câu hỏi
- Viết Individual Reflections
- Rút ra insights chính
- Đưa ra khuyến nghị tổng thể
- Viết hạn chế và hướng phát triển

5. Documentation

- README.md: Cả nhóm đóng góp
- TEAM_PLAN.md: Cả nhóm review
- Code comments: Từng người chịu trách nhiệm phần của mình

6. Quality Control

- Code review lẫn nhau
- Test notebooks
- Kiểm tra consistency
- Sửa lỗi và optimize

III. Timeline thực hiện

Giai đoạn 1: Chuẩn bị (Tuần 1-2)

- **Tuần 1:**
    - Tìm hiểu đề bài và yêu cầu
    - Lựa chọn dataset
    - Phân công công việc chi tiết
    - Setup môi trường làm việc (Git, Jupyter)
- **Tuần 2:**
    - Data Collection
    - EDA ban đầu

    - Xác định 6 câu hỏi nghiên cứu
    - Data Preprocessing

Giai đoạn 2: Phân tích (Tuần 3-6)

- **Tuần 3:**
    - **Vy:** Bắt đầu Q1
    - **Hậu:** Bắt đầu Q3
    - **Hoàng:** Bắt đầu Q5
- **Tuần 4:**
    - **Vy:** Hoàn thiện Q1, bắt đầu Q2
    - **Hậu:** Hoàn thiện Q3, bắt đầu Q4
    - **Hoàng:** Hoàn thiện Q5, bắt đầu Q6
- **Tuần 5:**
    - **Vy:** Hoàn thiện Q2
    - **Hậu:** Hoàn thiện Q4
    - **Hoàng:** Setup ML pipeline cho Q6
- **Tuần 6:**
    - **Hoàng:** Hoàn thiện Q6 (ML models, evaluation)
    - **Cả nhóm:** Review lẫn nhau, fix issues

Giai đoạn 3: Hoàn thiện (Tuần 7-8)

- **Tuần 7:**
    - Viết Conclusion Summary
    - Viết Individual Reflections
    - Hoàn thiện README.md
    - Tạo TEAM_PLAN.md
- **Tuần 8:**
    - **Hoàng:** Sửa lỗi Q5 (urban penalty)

- **Hoàng:** Thêm Statistical Explanation cho Q5
- **Hoàng:** Thêm Limitations cho Q5 và Q6
- **Vy & Hậu:** Update headers cho Q3, Q4
- **Cả nhóm:** Final review và polish

Giai đoạn 4: Nộp bài (Tuần 9)

- Điền thông tin cá nhân
- Run lại tất cả notebooks
- Hoàn thiện file TEAM PLAN AND WORK DISTRIBUTION.PDF
- Git commit & push
- Nộp bài

IV. Công cụ và môi trường làm việc

Phần mềm:

- **Python:** 3.8+
- **Jupyter Notebook/Lab:** Để viết và chạy notebooks
- **Anaconda:** Quản lý môi trường
- **Git:** Version control
- **VS Code:** Code editor

Thư viện Python:

pandas >= 1.3.0

numpy >= 1.21.0

matplotlib >= 3.4.0

seaborn >= 0.11.0

scikit-learn >= 1.0.0

xgboost >= 1.5.0

imbalanced-learn >= 0.9.0

scipy >= 1.7.0

Collaboration:

- **Git/GitHub:** Version control và collaboration
- **Google Drive/Docs:** Chia sẻ tài liệu
- **Zalo/Discord:** Communication
- **Google Meet:** Họp nhóm online

V. Quy trình làm việc

1. Version Control:

 # Mỗi thành viên làm việc trên branch riêng

git checkout -b feature/Q1-analysis *# Ví dụ cho Vy*

 # Commit thường xuyên

git add .

git commit -m "Add Q1 analysis and visualizations"

 # Push lên GitHub

git push origin feature/Q1-analysis

 # Tạo Pull Request để review

 # Sau khi review OK → Merge vào main

2. Code Review:

- Mỗi notebook phải được ít nhất 1 thành viên khác review
- Check: Code quality, logic, documentation, results
- Feedback qua GitHub Pull Request comments

3. Documentation:

- Code comments cho logic phức tạp
- Markdown cells giải thích từng bước

- README.md cập nhật liên tục
- Individual Reflections viết cuối cùng

4. Testing:

- Run lại notebook từ đầu (Restart & Run All)
- Verify kết quả
- Check không có lỗi
- Đảm bảo reproducibility

VI. Đóng góp của từng thành viên

Bùi Đoàn Thúy Vy:

- **Tỷ lệ đóng góp:** ~33%
- **Notebooks chính:** Q1, Q2
- **Công việc hỗ trợ:** EDA, Visualization design, Documentation
- **Phát hiện quan trọng:** Health-Lifestyle Paradox
- **Kỹ năng đóng góp:** Data visualization, Statistical analysis, Communication

Phạm Trần Trung Hậu:

- **Tỷ lệ đóng góp:** ~33%
- **Notebooks chính:** Q3, Q4
- **Công việc hỗ trợ:** Data preprocessing, Feature engineering, Testing
- **Phát hiện quan trọng:** Interaction effects (Smoking × Age, Glucose × Gender)
- **Kỹ năng đóng góp:** Statistical testing, Data preprocessing, Problem-solving

Nguyễn Việt Hoàng:

- **Tỷ lệ đóng góp:** ~34%
- **Notebooks chính:** Q5, Q6
- **Công việc hỗ trợ:** Machine Learning, Project management, Quality assurance
- **Phát hiện quan trọng:** Urban Penalty, Behavioral Personas
- **Đặc biệt:** thêm Statistical Explanation và Limitations

- **Kỹ năng đóng góp:** Machine Learning, Project coordination, Critical thinking

VII. Thách thức và giải pháp

Thách thức 1: Class Imbalance nghiêm trọng

- **Vấn đề:** 95% không đột quỵ, chỉ 5% có đột quỵ
- **Giải pháp:** Sử dụng SMOTE, class_weight, đánh giá bằng ROC-AUC/PR-AUC
- **Người xử lý:** Nguyễn Việt Hoàng (Q6)

Thách thức 2: Missing values

- **Vấn đề:** BMI 4%, smoking_status 30% thiếu
- **Giải pháp:** Median imputation cho BMI, giữ nguyên smoking_status như category
- **Người xử lý:** Phạm Trần Trung Hậu (Preprocessing)

Thách thức 3: Lỗi phân tích Q5

- **Vấn đề:** Kết luận sai "NỮ GIỚI" bị urban penalty nhiều hơn
- **Thực tế:** NAM GIỚI bị ảnh hưởng nhiều hơn (+6.7% vs +0.2%)
- **Giải pháp:** Kiểm tra lại data, sửa toàn bộ kết luận
- **Người xử lý:** Nguyễn Việt Hoàng

Thách thức 4: p > 0.05 trong Q5

- **Vấn đề:** Kết quả không có ý nghĩa thống kê
- **Giải pháp:** Viết Statistical Explanation, áp dụng Precautionary Principle
- **Người xử lý:** Nguyễn Việt Hoàng

Thách thức 5: Interaction effects phức tạp

- **Vấn đề:** Khó visualize và diễn giải 3-way interactions
- **Giải pháp:** Stratified analysis, faceted plots
- **Người xử lý:** Phạm Trần Trung Hậu (Q3, Q4)

Thách thức 6: Consistency giữa notebooks

- **Vấn đề:** Format, style, language không đồng nhất
- **Giải pháp:** Tạo style guide, regular code review, final polish

- **Người xử lý:** Cả nhóm, điều phối bởi Nguyễn Việt Hoàng

VIII. Bài học rút ra

Về kỹ thuật:

1. **Data quality là nền tảng** - "Garbage in, garbage out"
2. **Domain knowledge quan trọng như technical skills** - Cần hiểu y tế để phân tích đúng
3. **Visualization là chìa khóa** để communicate insights
4. **Statistical significance ≠ Clinical significance** - Học từ Q5
5. **Model interpretability** quan trọng trong medical domain

Về quy trình:

1. **Planning tiết kiệm thời gian** - Phân công rõ ràng từ đầu
2. **Regular sync-up** giúp catch issues sớm
3. **Code review improves quality** và học hỏi lẫn nhau
4. **Documentation** tiết kiệm thời gian cho cả nhóm
5. **Git workflow** cần thiết lập từ đầu

Về teamwork:

1. **Mỗi người có strengths khác nhau** - Cần leverage và complement
2. **Constructive feedback** giúp cả nhóm improve
3. **Flexibility** khi requirements thay đổi
4. **Communication is key** - Ask questions, clarify assumptions
5. **Celebrate small wins** để maintain motivation

IX. Kết quả đạt được

Về mặt kỹ thuật:

- 6 câu hỏi nghiên cứu phân tích đầy đủ, sâu sắc
- Format nhất quán toàn bộ notebooks
- Statistical tests và visualizations chất lượng cao

- ML với 3 models + Feature Importance
- Xử lý đúng class imbalance (SMOTE)
- Giải thích đầy đủ statistical vs clinical significance

Về nội dung:

- Phát hiện patterns quan trọng:
    - Urban Penalty (nam +6.7%, nữ +0.2%)
    - Health-Lifestyle Paradox
    - Interaction effects (Smoking × Age, Glucose × Gender)
- Limitations & Future Directions chi tiết
- Individual Reflections với phân công rõ ràng
- Khuyến nghị thực tiễn cho chính sách y tế

Về documentation:

- Markdown formatting nhất quán
- Code comments đầy đủ
- README và TEAM PLAN AND WORK DISTRIBUTIONhoàn chỉnh

X. Tài liệu tham khảo

Dataset:

- Kaggle - Stroke Prediction
  Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Tài liệu khoa học:

- WHO - Stroke Statistics: https://www.who.int/news-room/fact-sheets/detail/stroke
- CDC - Stroke Facts: https://www.cdc.gov/stroke/facts.htm
- Altman & Bland (1995) - Statistical vs Clinical Significance, BMJ

Thư viện Python:

- Pandas Documentation: https://pandas.pydata.org/
- Scikit-learn Documentation: https://scikit-learn.org/

- XGBoost Documentation: https://xgboost.readthedocs.io/
- Seaborn Gallery: https://seaborn.pydata.org/