Team Plan and Work Distribution

1. Executive summary

Objective: This project analyzes personal health indicators from the 2022 CDC BRFSS dataset to identify key risk factors for heart disease. The primary goals were to quantify the effects of multimorbidity, segment the population into meaningful risk profiles, and build interpretable machine learning models to predict heart attacks (`HadHeartAttack`).

Dataset: The analysis utilizes `heart_2022_with_nans.csv`, comprising approximately **445,000 rows** and **40 features**. Unlike pre-cleaned datasets, this version preserves missing values (approximately 44% of rows contain missing data), necessitating a robust handling strategy to avoid bias.

Core outcomes:

- **Identification of "Silent" risks:** 8.4% of heart attack patients presented with zero prior chronic conditions.
- **The "Buffer" effect:** Physical activity significantly mitigates the cardiovascular risks associated with poor sleep and mental distress.
- **Critical comorbidity window:** Patients with exactly two chronic conditions represent the highest volume risk group (~24% of cases).

---

2. Team & Responsibilities

Member	Key responsibilities	Percentage
Nguyễn Hữu Anh Trí	Analyzing numeric/categorical columns	100%
Nguyễn Hữu Anh Trí	Investigating multimorbidity (Question 2)	100%
Nguyễn Hữu Anh Trí	Building supervised risk prediction models (Question 4).	100%
Cao Tấn Hoàng Huy	Handling missing values and outliers	100%
Cao Tấn Hoàng Huy	Investigating the Mind–Body connection (Question 1)	100%
Cao Tấn Hoàng Huy	Performing patient clustering (Question 3).	100%

3. Data overview & preprocessing

3.1 Data dictionary & structure

The dataset consists of 40 columns, categorized into 7 distinct groups based on the nature of the indicators. The primary target variable for modeling is HadHeartAttack.

Group 1: Demographics & Basic information

- Key identifiers including `State`, `Sex` and `AgeCategory`.

- `RaceEthnicityCategory` is used to analyze disparities across demographic groups.

Group 2: Related Heart conditions (Target & History)

- `HadHeartAttack` (Target): Binary outcome indicating if the respondent ever had a heart attack.
- Includes related cardiovascular history: `HadAngina` (chest pain) and `HadStroke`.

Group 3: Physical health metrics

- General status: `GeneralHealth`, `PhysicalHealthDays` and `MentalHealthDays`.
- Biometrics: `HeightInMeters` and `WeightInKilograms`, which are used to derive `BMI` (Body Mass Index).

Group 4: Chronic conditions (Comorbidities)

- Includes major risk multipliers: `HadDiabetes`, `HadKidneyDisease`, `HadCOPD`, `HadAsthma`, `HadArthritis`, `HadDepressiveDisorder` and `HadSkinCancer`.

Group 5: Disabilities & Limitations

- Sensory issues: `DeafOrHardOfHearing`, `BlindOrVisionDifficulty`.
- Functional limitations: `DifficultyWalking`, `DifficultyConcentrating`, `DifficultyDressingBathing`, and `DifficultyErrands`.

Group 6: Lifestyle & Behaviors

- Substance use: `SmokerStatus`, `ECigaretteUsage` and `AlcoholDrinkers`.
- Habits: `PhysicalActivities` and `SleepHours`.

Group 7: Healthcare & Prevention

- Access and history: `LastCheckupTime`, `ChestScan`, `RemovedTeeth`.
- Vaccinations & Recent Risks: `FluVaxLast12`, `PneumoVaxEver`, `TetanusLast10Tdap`, `CovidPos`, `HighRiskLastYear` and `HIVTesting`.

3.2 Missing value handling strategy

A robust, multi-stage strategy was designed to handle missing data, prioritizing the preservation of physiological relationships over simple deletion.

Step 1: Strategic feature retention (The BMI Triangle)

Unlike standard cleaning pipelines that might drop redundant columns immediately, we **temporarily retained** `WeightInKilograms` and `HeightInMeters` alongside `BMI`.

- **Rationale:** These three variables are mathematically linked. By keeping all three, we can mathematically recover missing values with 100% accuracy if the other two are present, before resorting to statistical imputation.

Step 2: Dropping non-essential features

Columns identified in the initial audit as low-value were dropped **unless** they were part of the BMI triangle above.

- **Justification:** Features unrelated to cardiovascular health or those with excessive missingness (high missing-correlation ratio ~18%) were removed to reduce noise. This ensures the model focuses only on variables with proven clinical or predictive relevance.

Step 3: Tree-Based MICE imputation

For the remaining missing values that could not be mathematically recovered, we applied **Multivariate Imputation by Chained Equations (MICE)** using a **Tree-Based approach** (ExtraTrees Regressor) rather than standard Linear Regression.

- **Why Tree-Based?** Medical data is rarely linear; it is often clustered and threshold-based. A linear model assumes a straight-line relationship (e.g., "as age increases, risk increases constantly"), which fails to capture complex interactions.
- **Clinical advantage:** Tree-based models naturally capture non-linear, logical rules inherent in health data.
- *Example:* A linear imputer might average risk factors. A tree-based imputer can learn specific rules like: *"If Age > 60 AND Smoker = Yes, then ChestScan is likely Yes,"* providing a much more realistic imputation for complex patient profiles.

3.3 **Outlier management (Winsorization):**

To reduce noise while retaining data, extreme outliers were capped based on domain knowledge:

- `MentalHealthDays`: Capped at 30 days.
- `SleepHours`: Capped at 0.5 – 18 hours.
- `BMI`: Capped at 15 – 60.

Abnormal insight handling: Anomalies such as "Normal Weight" heart attacks or younger patients with high chronic disease rates were cross-verified. Inconsistent survey responses (e.g., Age mismatches) led to targeted row exclusion if unresolvable.

---

4. Exploratory data analysis

4.1 Numeric analysis

- **Distribution:** `BMI` and `MentalHealthDays` showed significant right-skewness. A log-transformation was applied for linear modeling, while tree-based models used the raw (Winsorized) data.
- **Correlations:** `GeneralHealth`, `PhysicalHealthDays`, and `AgeCategory` showed the strongest correlation with `HadHeartAttack`. Interestingly, `BMI` showed limited standalone predictive power,

suggesting it must be interpreted alongside Age and Sex.

4.2 Categorical analysis

- **Race & demographics:** Prevalence rates were calculated across racial groups. Preliminary analysis suggests disparities in heart disease prevalence, often compounded by socioeconomic factors implicit in the `State` data.
- **Simplification:** Granular categories for `SmokerStatus` and `HadDiabetes` (e.g., "Pre-diabetes") were grouped into binary flags to reduce noise and improve model stability.

---

5. Modeling & Key findings

Question 1: The Mind–Body connection

Hypothesis: Poor mental health and abnormal sleep increase heart risk, but physical activity can mitigate this.

- **Findings:**
- **U-Shaped sleep curve:** Both sleep deprivation (<6 hours) and oversleeping (>9 hours) are associated with higher heart attack risk.
- **Synergy:** High `MentalHealthDays` (distress) combined with poor sleep creates a compounding risk effect.
- **The protective buffer:** Interaction terms revealed that **Physical activity** acts as a moderator. Active individuals with poor sleep/mental health showed significantly lower risk compared to inactive individuals with the same stress levels.

Question 2: Compounding multimorbidity

Question: How does the accumulation of chronic conditions affect risk?

- **Method:** A **Comorbidity score (0–3+)** was calculated by summing chronic conditions (Diabetes, Kidney Disease, COPD, etc.).
- **Key Statistics:**
- **Score 0 (The silent threat): 8.4%** of heart attack sufferers had zero other chronic conditions, highlighting the need for screening beyond just "sick" patients.
- **Score 2 (The trap):** This group accounted for **~24%** of cases, representing the highest volume risk segment.
- **Age interaction:** The relative risk of multimorbidity is higher in younger age groups. While older adults have higher absolute risk, a young person with 2+ conditions faces a drastically higher odds ratio compared to their healthy peers.

Question 3: Patient profiling & clustering

Goal: Identify distinct patient archetypes using Unsupervised Learning.

- **Method:** K-Prototypes (for mixed numeric/categorical data) was used to cluster the population.
- **Identified profiles:**

1. **The "Older Multimorbid":** High age, multiple chronic conditions, sedentary. (Highest Risk).
2. **The "Young high-stress":** Younger demographic, physically capable, but high `MentalHealthDays` and poor sleep habits.
3. **The "Lifestyle risk":** Smokers and heavy drinkers who are otherwise "normal" weight and mobile.

- **Abnormal insight:** The "Lifestyle Risk" cluster proves that physical capability does not equal heart health; addiction plays a major role even in the absence of obesity.

Question 4: Risk prediction

Objective: Build an interpretable model to predict heart attacks.

- **Models:**
  - **Logistic Regression (ElasticNet):** Prioritized for interpretability.
  - **Random Forest:** Prioritized for high performance and capturing non-linear interactions.
  - **K-Nearest Neighbors**: Try to predict mainly base on dataset
- **Performance strategy:**
  - Given the class imbalance, **Recall (Sensitivity)** was the primary metric. It is clinically safer to flag a false positive than to miss a potential heart attack.
- **Feature Importance:** `AgeCategory`, `GeneralHealth`, and the custom `MultimorbidityScore` consistently ranked as top predictors. `BMI` was less predictive than expected.

---

6. Collaboration process & Workflow

The team adopted an agile, iterative data science pipeline to ensure efficiency and reproducibility. The collaboration was structured into three distinct phases:

Phase 1: Foundation & Pipeline design (Parallel execution) (30/11/2025 - 6/12/2025)

- **Role Division:** The team leveraged individual strengths to parallelize the initial workload.
- **Huy** took ownership of the **Data Engineering** pipeline, focusing on the complex cleaning strategy required for the "With NaNs" dataset. This involved researching and implementing MICE imputation to ensure the data was model-ready.
- **Trí** simultaneously initiated the **Exploratory Data Analysis (EDA)**, creating the statistical framework to understand the feature distributions (e.g., identifying the "BMI Triangle" relationship) and defining the hypothesis questions.

Phase 2: Integration & Hand-off (7/12/2025 - 13/12/2025)

- **The "Cleaned Data" Handshake:** A critical collaboration point was the generation of the intermediate file `cleaned_heart_data.csv`. Huy's cleaning script processed the raw CDC data, which was then validated and handed off to Trí. This ensured that both members were modeling on a unified, high-quality dataset, preventing consistency errors.
- **Code Review:** The team conducted cross-checks on critical decisions, such as the choice to remove duplicate rows (validated by the high dimensionality argument) and the decision to cap outliers rather

than drop them.

Phase 3: Modeling & Synthesis (14/12/2025 - 20/12/2025)

- **Modular Modeling:** To answer the four core research questions efficiently, the modeling tasks were modularized:
- Huy focused on **Unsupervised Learning** (Clustering/Patient Profiling) and the "Mind-Body" interaction (Q1).
- Trí focused on **Supervised Learning** (Risk Prediction) and Multimorbidity analysis (Q2).
- **Insight Synthesis:** Final insights were synthesized by comparing results. For instance, Trí's finding that "BMI is a weak predictor" was contextualized by Huy's cluster analysis, which found a "High Risk" cluster that was not necessarily obese, leading to the unified conclusion that *lifestyle factors often outweigh simple biometrics*.

Phase 4: Refine reports and markdowns analysis (21/12/2025 - 27/12/2025)

- Both re-check the analysis and finish reports, readme files for the projects.

Tools & Technologies

- **Version Control & Sharing:** The project utilized **Jupyter Notebooks** for code development, with regular synchronization to merge the *Exploration* and *Modeling* modules.
- **Python Stack:** The team standardized on a shared library stack (`pandas`, `scikit-learn`, `matplotlib`, `seaborn`) to ensure code compatibility and reproducibility across different machines.
