


Giai đoạn	Công việc chi tiết	KPI/Output cần sinh	Thành viên phụ trách	Công cụ & deliverable
1. Hiểu & thống kê sơ bộ	<ul style="list-style-type: none"> - Đọc mô tả trường dữ liệu, xác định kiểu & ý nghĩa - Khảo sát % thiếu giá trị, giá trị ngoại lệ (box-plot, z-score, IQR) - Thống kê mô tả (mean, median, std...) cho biến số & phân bố cho biến hạng 	<ul style="list-style-type: none"> • Data Dictionary chuẩn hóa • Báo cáo quality issues (NA, outlier) 	Don	Jupyter + Pandas; 1 notebook EDA_raw.ipynb
2. Làm sạch & chuẩn hoá	<ul style="list-style-type: none"> - Xử lý NA (bỏ, suy diễn, median impute) - Chuẩn hóa tên cột, kiểu dữ liệu (datetime, category) - Gộp/nhóm age, income thành band - Sinh biến mới: TenureBand, IncomeBand, WorkLifeBalanceScore (mức 1-4) 	<ul style="list-style-type: none"> • Dataset ibm_hr_clean.csv • Log cleaning chi tiết 	Don	Pandas + Great Expectations; Notebook cleaning.ipynb
3. Khám phá trực quan (EDA)	<ul style="list-style-type: none"> - Biểu đồ phân bố Attrition overall và theo: <ul style="list-style-type: none"> • Department, Job Role • Age Band, TenureBand • Overtime, JobSatisfaction, EnvSatisfaction - Heatmap tương quan biến số - Sankey flow “đi làm rồi rời đi” theo hành trình 	<ul style="list-style-type: none"> • 10-12 figure PNG/SVG (đặt tên rõ ràng) • Insight mô tả (bullet) 	Hùng	Matplotlib/Seaborn, Plotly; Notebook eda_visual.ipynb

Giai đoạn	Công việc chi tiết	KPI/Output cần sinh	Thành viên phụ trách	Công cụ & deliverable
4. KPI cốt lõi	1. Attrition Rate toàn công ty = số nghỉ / tổng 2. Attrition Rate theo Department, Job Role 3. Attrition Rate theo Age Band & TenureBand 4. Average Monthly Income theo nhóm ở lại vs. nghỉ 5. % Overtime nhân viên nghỉ vs. ở lại 6. Satisfaction Gap (Job / Environment) 7. Cost of Turnover $\approx 1.5 \times \text{MonthlyIncome} \times \text{người nghỉ}$	<ul style="list-style-type: none"> Bảng KPI (csv + markdown) Dashboard (interactive) 	Phần KPI này chia bên dưới	Pandas + Plotly Dash (hoặc Power BI)
5. Phân tích nguyên nhân & mô hình dự đoán	- Tiền xử lý: one-hot, scale - Chia train/test, oversample nếu cần - Xây Decision Tree và Logistic Regression - Đo độ chính xác, ROC-AUC, feature importance - Xuất rules quan trọng (Overtime \rightarrow Attrition...)	<ul style="list-style-type: none"> Model report (accuracy, AUC, confusion matrix) Top 10 feature importance Cây quyết định hình ảnh 	Đôn, Hùng	Scikit-learn; Notebook modeling.ipynb
6. Data Storytelling & báo cáo	- Viết kịch bản theo cấu trúc 4 phần bạn đã phác thảo (mở đầu, cao trào, tìm nguyên nhân, hành động) - Chọn 6-8 visual then chèn vào slide	<ul style="list-style-type: none"> Bộ slide 15-18 trang (PPTX) Script thuyết trình (Markdown) 	Hải	PowerPoint / Canva; File storytelling.pptx , script.md

Giai đoạn	Công việc chi tiết	KPI/Output cần sinh	Thành viên phụ trách	Công cụ & deliverable
	<ul style="list-style-type: none"> - Thiết kế slide theo quy tắc <i>Less is More</i>, màu thống nhất - Viết chú giải (speaker notes) chứa thông điệp dữ liệu 			
7. Kiểm thử & hoàn thiện	<ul style="list-style-type: none"> - Peer-review số liệu & hình ảnh (swap file kiểm tra) - Chạy lại notebook end-to-end, lưu HTML report - Chuẩn bị câu hỏi dự kiến & response 	<ul style="list-style-type: none"> • Final repo ready to deliver 	Cả 3	Git + README với hướng dẫn chạy
8. Triển khai & trình bày	<ul style="list-style-type: none"> - Don demo pipeline cleaning-EDA-model (5') - Member 2 trình bày kết quả & KPI (7') - Member 3 kể chuyện & khuyến nghị (8') 	<ul style="list-style-type: none"> • Buổi trình bày 20' + 10' Q&A 	Cả 3	Slide deck + live demo

KPI	Có hỗ trợ "storytelling" mượt mà không?	Vì sao?	Vị trí phù hợp trong câu chuyện
Attrition Rate theo Department/JobRole	 Rất tốt	Phát hiện ra “ổ dịch” – nơi vấn đề rõ nhất, tạo cao trào.	Cao trào (Đôn)

KPI	Có hỗ trợ "storytelling" mượt mà không?	Vì sao?	Vị trí phù hợp trong câu chuyện
Attrition Rate theo Age Band/TenureBand	✅ Rất tốt	Hé lộ nhóm nhân sự dễ rời đi – tiếp tục đẩy cao trào.	Cao trào (Hùng)
% Overtime nghỉ vs. ở lại	✅ Rất tốt	Chỉ rõ một nguyên nhân sâu xa cực kỳ thuyết phục.	Nguyên nhân (Hải)
Satisfaction Gap	✅ Tốt	Bổ sung yếu tố cảm xúc rất hay vào câu chuyện (không chỉ số liệu lạnh lùng).	Nguyên nhân (Hải)
Average Monthly Income nghỉ vs. ở lại	✅ Tốt	Cho thêm một nguyên nhân khác về tài chính (money matters).	Nguyên nhân (Hải)

Biến	Kiểu dữ liệu	Ý nghĩa	Ghi chú phân tích
Age	numeric (int)	Tuổi nhân viên	Có thể chia thành nhóm AgeBand
Attrition	category (Yes/No)	Có nghỉ việc không	Biến mục tiêu (Target Variable)
BusinessTravel	category	Tần suất đi công tác	Có 3 mức: Rarely, Frequently, Non-Travel
DailyRate	numeric (int)	Mức lương ngày (không quan trọng)	Không dùng mạnh trong phân tích chính
Department	category	Bộ phận làm việc	Sales, Research & Development, HR

Biến	Kiểu dữ liệu	Ý nghĩa	Ghi chú phân tích
DistanceFromHome	numeric (int)	Khoảng cách từ nhà đến công ty (mile)	Ảnh hưởng một phần đến turnover
Education	ordinal (1-5)	Trình độ học vấn (1: thấp nhất, 5: cao nhất)	Có thể kết hợp thêm EducationField
EducationField	category	Ngành học	Life Sciences, Technical Degree, Human Resources...
EmployeeCount	numeric	Số lượng nhân viên	Luôn = 1 (vô nghĩa, bỏ)
EmployeeNumber	numeric (ID)	ID nhân viên	Chỉ để định danh, không phân tích
EnvironmentSatisfaction	ordinal (1-4)	Mức độ hài lòng với môi trường làm việc	Rất quan trọng để phân tích nguyên nhân
Gender	category	Nam/Nữ	Không phải yếu tố chính nhưng có thể tham khảo
HourlyRate	numeric	Lương theo giờ	Thường bỏ, dùng MonthlyIncome tốt hơn
JobInvolvement	ordinal (1-4)	Mức độ tham gia công việc	Thể hiện mức độ gắn kết
JobLevel	ordinal	Cấp bậc công việc (1 thấp → 5 cao)	Tương quan với lương
JobRole	category	Chức danh công việc	Sales Executive, Research Scientist,...
JobSatisfaction	ordinal (1-4)	Mức độ hài lòng công việc	Rất quan trọng để phân tích nguyên nhân
MaritalStatus	category	Tình trạng hôn nhân	Single, Married, Divorced

Biến	Kiểu dữ liệu	Ý nghĩa	Ghi chú phân tích
MonthlyIncome	numeric	Tổng lương hàng tháng	Một trong các biến nguyên nhân chính
MonthlyRate	numeric	Mức lương hàng tháng khác (không cần thiết)	Bỏ, gây nhiễu
NumCompaniesWorked	numeric	Số công ty từng làm	Thể hiện mức độ nhảy việc
Over18	category	Tất cả đều Y	Không có ý nghĩa phân tích, bỏ luôn
OverTime	category	Có làm thêm giờ không	Cực kỳ quan trọng – yếu tố stress!
PercentSalaryHike	numeric	% tăng lương gần nhất	Thể hiện đãi ngộ
PerformanceRating	ordinal	Đánh giá hiệu suất (thường cao đều)	Ít phân biệt rõ ràng
RelationshipSatisfaction	ordinal	Mức độ hài lòng về mối quan hệ tại nơi làm việc	Bổ sung insight
StandardHours	numeric	Tất cả = 80	Không có ý nghĩa phân tích, bỏ luôn
StockOptionLevel	ordinal (0–3)	Quyền chọn cổ phiếu	Quyền lợi tài chính
TotalWorkingYears	numeric	Tổng số năm đi làm	Gắn với thâm niên
TrainingTimesLastYear	numeric	Số lần đào tạo trong năm	Phát triển nghề nghiệp
WorkLifeBalance	ordinal (1–4)	Cân bằng công việc-cuộc sống	Chỉ số rất quan trọng
YearsAtCompany	numeric	Số năm làm ở công ty hiện tại	Sẽ gộp thành TenureBand

Biến	Kiểu dữ liệu	Ý nghĩa	Ghi chú phân tích
YearsInCurrentRole	numeric	Số năm giữ chức vụ hiện tại	
YearsSinceLastPromotion	numeric	Số năm từ lần thăng chức gần nhất	
YearsWithCurrManager	numeric	Số năm làm với quản lý hiện tại	Mối quan hệ quản lý