

ĐỀ CƯƠNG ÔN TẬP

KHAI PHÁ DỮ LIỆU

Dành cho sinh viên đại học khối kinh tế, tài chính, công nghệ tài chính

Data Mining — a bilingual study guide

Dữ liệu và tiền xử lý · Luật kết hợp · Cây quyết định, Naive Bayes

kNN, hồi quy, mạng nơ-ron · Đánh giá mô hình

Phân cụm · Phát hiện bất thường, khai phá văn bản · Pháp luật về dữ liệu

Bài tập tính tay có lời giải · Tự luận có dàn ý · Trắc nghiệm có đáp án · Mã Python chạy thật

Song ngữ Việt – Anh · Vietnamese – English

Phiên bản 1.0 — Tháng 10 năm 2026

<https://globalmarket.vn>

M U C L U C

CONTENTS

Chương 1. Tổng quan về khai phá dữ liệu — Chapter 1. Introduction to data mining ..	6
1.1. Dữ liệu, thông tin và tri thức — 1.1. Data, information and knowledge	6
1.2. Khai phá dữ liệu và quá trình KDD — 1.2. Data mining and the KDD process	10
1.3. CRISP-DM và dự án chấm điểm tín dụng của Phú Sa — 1.3. CRISP-DM and Phú Sa's credit scoring project	13
1.4. Nhiệm vụ mô tả, nhiệm vụ dự đoán và các bài toán khai phá — 1.4. Descriptive and predictive tasks and the mining problems	16
1.5. Học có giám sát, không giám sát và bán giám sát — 1.5. Supervised, unsupervised and semi-supervised learning.....	19
1.6. Các loại dữ liệu được khai phá — 1.6. Types of data that are mined	23
1.7. Kho dữ liệu, OLAP và sự khác biệt với khai phá dữ liệu — 1.7. Data warehouse, OLAP and how they differ from data mining	26
1.8. Ứng dụng trong tài chính và công cụ khai phá dữ liệu — 1.8. Applications in finance and data mining tools	28
1.9. Quy trình dự án, vai trò nhân sự và thách thức — 1.9. Project process, team roles and challenges	31
1.10. Khung pháp lý tóm lược — 1.10. A short legal framework.....	33
1.11. Bảng thuật ngữ — 1.11. Glossary	35
1.12. Bài tập vận dụng có lời giải — Applied exercises with solutions	38
1.13. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines	40
1.14. Câu hỏi trắc nghiệm — Multiple-choice questions	42
Chương 2. Dữ liệu, thống kê mô tả và độ đo tương tự — Chapter 2. Data, descriptive statistics and similarity measures.....	44
2.1. Đối tượng, thuộc tính và kiểu thuộc tính — 2.1. Objects, attributes and attribute types	44
2.2. Đo xu hướng trung tâm — 2.2. Measures of central tendency	48
2.3. Đo độ phân tán và năm số tóm tắt — 2.3. Measures of dispersion and the five-number summary	51
2.4. Hình hộp, histogram và độ lệch — 2.4. Boxplots, histograms and skewness	55
2.5. Tương quan Pearson và biểu đồ phân tán — 2.5. Pearson correlation and the scatter plot.....	57
2.6. Ma trận dữ liệu, ma trận khoảng cách và khoảng cách Minkowski — 2.6. Data matrix, dissimilarity matrix and Minkowski distance.....	59
2.7. Độ tương tự của thuộc tính nhị phân: SMC và Jaccard — 2.7. Similarity of binary attributes: SMC and Jaccard	65
2.8. Thuộc tính định danh, thứ bậc và độ tương tự cosine — 2.8. Nominal and ordinal attributes and cosine similarity	67
2.9. Thuộc tính hỗn hợp và khoảng cách Gower — 2.9. Mixed attributes and the Gower distance.....	69

2.10. Bảng thuật ngữ — 2.10. Glossary	72
2.11. Bài tập vận dụng có lời giải — Applied exercises with solutions	74
2.12. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines	77
2.13. Câu hỏi trắc nghiệm — Multiple-choice questions	79
Chương 3. Tiền xử lý dữ liệu — Chapter 3. Data preprocessing	81
3.1. Chất lượng dữ liệu và các bước tiền xử lý — 3.1. Data quality and the preprocessing steps	81
3.2. Làm sạch: giá trị thiếu — 3.2. Cleaning: missing values	84
3.3. Làm sạch: nhiễu, chia giỏ và làm trơn — 3.3. Cleaning: noise, binning and smoothing	86
3.4. Ngoại lai: phát hiện theo IQR và z-score — 3.4. Outliers: detection by IQR and z-score	89
3.5. Tích hợp dữ liệu — 3.5. Data integration	92
3.6. Biến đổi: chuẩn hóa dữ liệu — 3.6. Transformation: normalisation	94
3.7. Biến đổi: mã hóa, rời rạc hóa và thuộc tính mới — 3.7. Transformation: encoding, discretisation and new attributes	97
3.8. Thu giảm số chiều: chọn thuộc tính và PCA — 3.8. Dimensionality reduction: attribute selection and PCA	100
3.9. Lấy mẫu và mất cân bằng lớp — 3.9. Sampling and class imbalance	104
3.10. Rò rỉ dữ liệu và thứ tự đúng của các bước — 3.10. Data leakage and the correct order of steps	106
3.11. Bảng thuật ngữ — 3.11. Glossary	108
3.12. Bài tập vận dụng có lời giải — Applied exercises with solutions	109
3.13. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines	111
3.14. Câu hỏi trắc nghiệm — Multiple-choice questions	113
Chương 4. Khai phá luật kết hợp — Chapter 4. Association rule mining	115
4.1. Bài toán phân tích giỏ hàng và bán chéo sản phẩm tài chính — 4.1. Market basket analysis and cross-selling of financial products	115
4.2. Tập mục, độ hỗ trợ, độ tin cậy, lift và các độ đo khác — 4.2. Itemsets, support, confidence, lift and other measures	118
4.3. Thuật toán Apriori: tính chất đóng xuống, ghép và tỉa — 4.3. The Apriori algorithm: downward closure, join and prune	121
4.4. Bài mẫu 1: chạy tay Apriori trên giao dịch sản phẩm của Phú Sa — 4.4. Worked example 1: Apriori by hand on the Phú Sa product transactions	124
4.5. Sinh luật từ tập phổ biến và lọc theo độ tin cậy — 4.5. Generating rules from frequent itemsets and filtering by confidence	127
4.6. Luật mạnh nhưng không thú vị: đánh giá bằng lift, leverage và conviction — 4.6. Strong but uninteresting rules: evaluation with lift, leverage and conviction	130
4.7. Tập phổ biến đóng và tập phổ biến cực đại — 4.7. Closed and maximal frequent itemsets	132
4.8. Thuật toán FP-growth và cây FP-tree — 4.8. The FP-growth algorithm and the FP-tree	134

4.9. So sánh Apriori và FP-growth; luật nhiều mức và luật định lượng — 4.9. <i>Apriori compared with FP-growth; multilevel and quantitative rules</i>	139
4.10. Mẫu tuần tự, ứng dụng đề xuất sản phẩm trên ứng dụng Phú Sa và giới hạn — 4.10. <i>Sequential patterns, product recommendation in the Phú Sa app and limitations</i> ..	142
4.11. Bảng thuật ngữ — 4.11. <i>Glossary</i>	144
4.12. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	146
4.13. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	149
4.14. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	151

Chương 5. Phân lớp: cây quyết định và Naive Bayes — Chapter 5. Classification: decision trees and Naive Bayes 153

5.1. Bài toán phân lớp và dữ liệu hồ sơ vay — 5.1. <i>The classification problem and the loan-application data</i>	153
5.2. Cây quyết định: cấu trúc và quy nạp từ trên xuống — 5.2. <i>Decision trees: structure and top-down induction</i>	156
5.3. Entropy và information gain: tiêu chí của ID3 — 5.3. <i>Entropy and information gain: the ID3 criterion</i>	158
5.4. Gain ratio, chỉ số Gini và phép tách nhị phân — 5.4. <i>Gain ratio, the Gini index and binary splits</i>	160
5.5. Dựng cây ID3 hoàn chỉnh và rút luật IF–THEN — 5.5. <i>Building the full ID3 tree and extracting IF–THEN rules</i>	165
5.6. Thuộc tính liên tục và ngưỡng tách — 5.6. <i>Continuous attributes and split thresholds</i>	169
5.7. Quá khớp, tia cây, ưu điểm và nhược điểm — 5.7. <i>Overfitting, pruning, strengths and weaknesses</i>	171
5.8. Naive Bayes: định lý Bayes, làm trơn Laplace và log-xác suất — 5.8. <i>Naive Bayes: Bayes' theorem, Laplace smoothing and log-probabilities</i>	174
5.9. Gaussian Naive Bayes và so sánh với cây quyết định — 5.9. <i>Gaussian Naive Bayes and comparison with decision trees</i>	178
5.10. Bảng thuật ngữ — 5.10. <i>Glossary</i>	182
5.11. Bài tập vận dụng có lời giải — <i>Applied exercises with solutions</i>	184
5.12. Câu hỏi tự luận và dàn ý trả lời — <i>Essay questions and answer outlines</i>	187
5.13. Câu hỏi trắc nghiệm — <i>Multiple-choice questions</i>	189

Chương 6. Phân lớp nâng cao và mô hình dự báo — Chapter 6. Advanced classification and predictive models 191

6.1. Hai bài toán dự đoán và bộ hồ sơ thử nghiệm — 6.1. <i>Two prediction problems and the test applications</i>	191
6.2. k láng giềng gần nhất (kNN) — 6.2. <i>k-nearest neighbours (kNN)</i>	193
6.3. Hồi quy tuyến tính dự báo dư nợ — 6.3. <i>Linear regression to forecast outstanding loans</i>	198
6.4. Hồi quy logistic — 6.4. <i>Logistic regression</i>	202
6.5. Perceptron, mạng nơ-ron và học sâu — 6.5. <i>The perceptron, neural networks and deep learning</i>	204
6.6. Máy véc-tơ hỗ trợ (SVM) — 6.6. <i>Support vector machines (SVM)</i>	209

6.7. Học kết hợp: bagging, rừng ngẫu nhiên, boosting — 6.7. Ensemble learning: bagging, random forests, boosting	212
6.8. So sánh các bộ phân lớp và chọn mô hình — 6.8. Comparing classifiers and choosing a model	216
6.9. Bảng thuật ngữ — 6.9. Glossary	218
6.10. Bài tập vận dụng có lời giải — Applied exercises with solutions	221
6.11. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines	223
6.12. Câu hỏi trắc nghiệm — Multiple-choice questions	225

Chương 7. Đánh giá và lựa chọn mô hình — Chapter 7. Model evaluation and selection

7.1. Ma trận nhầm lẫn và bộ dữ liệu điểm dự báo — 7.1. The confusion matrix and the predicted-score dataset	227
7.2. Accuracy, precision, recall, specificity, F1 và F β — 7.2. Accuracy, precision, recall, specificity, F1 and F β	229
7.3. Ngưỡng quyết định và đánh đổi precision – recall — 7.3. The decision threshold and the precision–recall trade-off	232
7.4. Đường ROC, AUC và Gini — 7.4. The ROC curve, AUC and the Gini coefficient	235
7.5. Đường precision – recall, gain và lift — 7.5. The precision–recall curve, gain and lift	238
7.6. Ma trận chi phí và chọn ngưỡng — 7.6. The cost matrix and threshold selection	240
7.7. Ước lượng sai số: holdout, k-fold, leave-one-out, bootstrap — 7.7. Estimating error: holdout, k-fold, leave-one-out, bootstrap	243
7.8. Quá khớp, chưa khớp và bias – variance — 7.8. Overfitting, underfitting and bias–variance	247
7.9. So sánh hai mô hình bằng khoảng tin cậy — 7.9. Comparing two models with a confidence interval	250
7.10. Chỉ số hồi quy và đánh giá trong chấm điểm tín dụng — 7.10. Regression measures and evaluation in credit scoring	255
7.11. Bảng thuật ngữ — 7.11. Glossary	257
7.12. Bài tập vận dụng có lời giải — Applied exercises with solutions	259
7.13. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines	261
7.14. Câu hỏi trắc nghiệm — Multiple-choice questions	263

Chương 8. Phân tích cụm — Chapter 8. Cluster analysis

8.1. Khái niệm, yêu cầu và bài toán phân cụm — 8.1. Concept, requirements and the clustering task	265
8.2. Phân loại các phương pháp phân cụm — 8.2. Families of clustering methods	267
8.3. Chuẩn hóa dữ liệu trước khi phân cụm — 8.3. Scaling the data before clustering	269
8.4. Phân cụm phân hoạch: thuật toán k-means — 8.4. Partitioning: the k-means algorithm	270
8.5. Chọn số cụm k và đánh giá chất lượng cụm — 8.5. Choosing k and judging cluster quality	275
8.6. Nhược điểm của k-means, k-means++ và k-medoids — 8.6. Weaknesses of k-means, k-means++ and k-medoids	279

8.7. Phân cụm phân cấp — 8.7. Hierarchical clustering	283
8.8. Phân cụm theo mật độ: DBSCAN — 8.8. Density-based clustering: DBSCAN.....	287
8.9. Phân khúc khách hàng bằng RFM — 8.9. Customer segmentation with RFM.....	291
8.10. Bảng thuật ngữ — 8.10. Glossary	294
8.11. Bài tập vận dụng có lời giải — Applied exercises with solutions	296
8.12. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines	298
8.13. Câu hỏi trắc nghiệm — Multiple-choice questions	300
Chương 9. Phát hiện bất thường, khai phá văn bản và các vấn đề pháp lý – đạo đức — Chapter 9. Anomaly detection, text mining, and legal and ethical issues	303
9.1. Bất thường và ngoại lai — 9.1. Anomalies and outliers	303
9.2. Phương pháp thống kê: z-score, IQR và MAD — 9.2. Statistical methods: z-score, IQR and MAD.....	305
9.3. Phương pháp dựa khoảng cách, mật độ và mô hình — 9.3. Distance-, density- and model-based methods	309
9.4. Ứng dụng chống gian lận thẻ và phòng chống rửa tiền — 9.4. Applications to card fraud and anti-money-laundering	313
9.5. Chuỗi thời gian: thành phần, trung bình trượt và san mũ — 9.5. Time series: components, moving average and exponential smoothing.....	315
9.6. Khai phá văn bản: tách từ, TF-IDF và độ tương tự cosin — 9.6. Text mining: word segmentation, TF-IDF and cosine similarity	320
9.7. Dữ liệu lớn và mô hình MapReduce — 9.7. Big data and the MapReduce model...	325
9.8. Khung pháp lý về dữ liệu và bí mật thông tin khách hàng — 9.8. Legal framework for data and customer-information secrecy.....	327
9.9. Trí tuệ nhân tạo và chấm điểm tự động — 9.9. Artificial intelligence and automated scoring.....	331
9.10. Đạo đức: thiên lệch, giải thích được, ẩn danh hóa và k-anonymity — 9.10. Ethics: bias, explainability, anonymisation and k-anonymity.....	334
9.11. Bảng thuật ngữ — 9.11. Glossary	339
9.12. Bài tập vận dụng có lời giải — Applied exercises with solutions	341
9.13. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines	344
9.14. Câu hỏi trắc nghiệm — Multiple-choice questions	346
Phụ lục A. Đáp án câu hỏi trắc nghiệm — Appendix A. Answer key.....	349
Phụ lục B. Công thức khai phá dữ liệu — Appendix B. Data mining formulas.....	354
Phụ lục C. Thuật ngữ Việt – Anh — Appendix C. Vietnamese–English glossary.....	357
Phụ lục D. Văn bản pháp luật về dữ liệu và trí tuệ nhân tạo — Appendix D. Vietnamese legislation on data and artificial intelligence	365
Tài liệu tham khảo — References.....	366

Chương 1. Tổng quan về khai phá dữ liệu — Chapter 1. Introduction to data mining

Mục tiêu chương — Chapter objectives

- Phân biệt dữ liệu, thông tin và tri thức; phân biệt KDD với khai phá dữ liệu và nêu năm bước của quá trình.
Distinguish data, information and knowledge; distinguish KDD from data mining and state the five steps of the process.
- Mô tả sáu pha CRISP-DM và áp vào dự án chấm điểm tín dụng của Công ty Tài chính TNHH Phú Sa.
Describe the six CRISP-DM phases and apply them to the credit scoring project of Phú Sa Finance Company Limited.
- Xếp một câu hỏi kinh doanh vào bài toán khai phá (phân lớp, hồi quy, phân cụm, luật kết hợp, bất thường, chuỗi, văn bản) và kiểu học tương ứng.
Place a business question under a mining problem (classification, regression, clustering, association rules, anomalies, sequences, text) and the matching learning type.
- Đọc bảng dữ liệu nhỏ, đếm, tính tỷ lệ và chấm một quy tắc phân lớp đơn giản bằng tay và bằng mã Python.
Read a small data table, count, compute rates and score a simple classification rule by hand and in Python code.
- Phân biệt kho dữ liệu, OLAP với khai phá dữ liệu; nêu thách thức của dự án và các quy định về dữ liệu cá nhân, bí mật thông tin khách hàng.
Tell a data warehouse and OLAP from data mining; state the challenges of a project and the rules on personal data and customer information confidentiality.

1.1. Dữ liệu, thông tin và tri thức — 1.1. Data, information and knowledge

Dữ liệu (data) là giá trị ghi lại về một đối tượng hay một sự kiện, chưa gắn với câu hỏi nào. Dòng "KH03, 34, Giáo viên, Vàng" trong bảng khách hàng của Công ty Tài chính TNHH Phú Sa (TP. Cần Thơ) là dữ liệu. Thông tin (information) là dữ liệu đã được tổ chức và đặt vào ngữ cảnh để trả lời một câu hỏi cụ thể, chẳng hạn "12 khách hàng, trong đó 7 người có nhà". Tri thức (knowledge) là mẫu hay quy luật rút ra từ nhiều thông tin, dùng được để quyết định cho trường hợp chưa gặp, chẳng hạn "hồ sơ có lịch sử tín dụng Tốt hiếm khi thành nợ xấu". Ba cấp này là phần dưới của mô hình DIKW (data, information, knowledge, wisdom). Khai phá dữ liệu dùng thuật toán để đi từ cấp thứ nhất lên cấp thứ ba.

Data is a recorded value about an object or an event, not yet tied to any question. The row "KH03, 34, Teacher, Gold" in the customer table of Phú Sa Finance Company Limited (Cần Thơ City) is data. Information is data that has been organised and placed in context to answer a specific question, for example "12 customers, 7 of whom own a house". Knowledge is a pattern or rule drawn from many pieces of information and usable for a decision on a case not yet seen, for example "applications with a Good credit history rarely become bad debt". The three levels

form the lower part of the DIKW model (data, information, knowledge, wisdom). Data mining uses algorithms to climb from the first level to the third.

Cấp Level	Đặc điểm Feature	Ví dụ ở Phú Sa Example at Phú Sa
Dữ liệu Data	Giá trị thô, chưa có ngữ cảnh, đọc một mình chưa trả lời được câu hỏi nào <i>Raw value without context; read alone it answers no question</i>	Dòng của KH07: 45 tuổi, kinh doanh tự do, thiếu thu nhập <i>The row of KH07: aged 45, self-employed, income missing</i>
Thông tin Information	Dữ liệu đã đếm, tổng hợp, so sánh để trả lời câu hỏi "bao nhiêu, ở đâu, khi nào" <i>Data counted, summarised and compared to answer "how many, where, when"</i>	7 trong 12 khách có nhà (58,3%) <i>7 of 12 customers own a house (58.3%)</i>
Tri thức Knowledge	Mẫu, quy luật hoặc mô hình khái quát được cho dữ liệu mới, dùng để quyết định <i>A pattern, rule or model that generalises to new data and supports a decision</i>	Lịch sử tín dụng Xấu đi cùng tỷ lệ nợ xấu 80%, lịch sử Tốt đi cùng 0% <i>A Bad credit history goes with an 80% bad-debt rate, a Good one with 0%</i>

Mã sau đọc 12 dòng bảng khách hàng (dữ kiện đề bài rút gọn còn năm cột) và đếm. Mỗi dòng in ra là một mẫu thông tin.

The following code reads the 12 rows of the customer table (a problem statement reduced to five columns) and counts. Each printed line is one piece of information.

```
from collections import Counter

khach = [ # (mã, tuổi, thu nhập triệu/tháng, nghề, có nhà)
    ("KH01", 25, 9, "Công nhân", 0),
    ("KH02", 29, 12, "Nhân viên văn phòng", 0),
    ("KH03", 34, 15, "Giáo viên", 1),
    ("KH04", 41, 22, "Kinh doanh tự do", 1),
    ("KH05", 23, 8, "Công nhân", 0),
    ("KH06", 38, 18, "Nhân viên văn phòng", 1),
    ("KH07", 45, None, "Kinh doanh tự do", 1),
    ("KH08", 31, 14, "Giáo viên", 0),
    ("KH09", 27, 11, "Nhân viên văn phòng", 0),
    ("KH10", 36, 16, "Công nhân", 1),
    ("KH11", 50, 20, "Giáo viên", 1),
    ("KH12", 47, 120, "Kinh doanh tự do", 1),
]
n = len(khach)
print("Số khách:", n)
print("Thiếu thu nhập:", [k[0] for k in khach if k[2] is None])
for nghe, c in Counter(k[3] for k in khach).items():
    print(f" {nghe}: {c} khách")
co_nha = sum(k[4] for k in khach)
print(f"Có nhà: {co_nha}/{n} = {co_nha / n:.1%}")
tuoi = Counter("<30" if k[1] < 30 else "30-39" if k[1] < 40 else ">=40"
               for k in khach)
print("Nhóm tuổi:", dict(tuoi))
```

Kết quả chạy / Output

```
Số khách: 12
Thiếu thu nhập: ['KH07']
Công nhân: 3 khách
Nhân viên văn phòng: 3 khách
Giáo viên: 3 khách
Kinh doanh tự do: 3 khách
Có nhà: 7/12 = 58.3%
Nhóm tuổi: {'<30': 4, '30-39': 4, '>=40': 4}
```

Bảng có 12 khách, một ô thiếu (thu nhập của KH07), mỗi nghề 3 khách, 7 trên 12 khách (58,3%) có nhà, mỗi nhóm tuổi 4 khách. Các con số mô tả bảng nhưng chưa giải thích vì sao khách này có nhà còn khách kia thì không; đó là việc của các bước sau.

The table has 12 customers, one missing cell (the income of KH07), 3 customers in each occupation, 7 of 12 customers (58.3%) owning a house, and 4 customers in each age group. The figures describe the table but do not explain why one customer owns a house and another does not; that is the job of the later steps.

Bài mẫu 1. Từ thông tin đến quy tắc — *Worked example 1. From information to a rule*

Đề bài: bảng 14 hồ sơ vay gồm lịch sử tín dụng (Tốt, Xấu, Chưa có) và nhãn nợ xấu (5 hồ sơ có nợ xấu). Tính tỷ lệ nợ xấu từng nhóm, đề xuất quy tắc phân lớp dùng một thuộc tính, tính độ chính xác và so với cách luôn đoán lớp đông nhất.

Problem: a table of 14 loan applications with credit history (Good, Bad, None) and a bad-debt label (5 applications with bad debt). Compute the bad-debt rate of each group, propose a classification rule using one attribute, compute its accuracy and compare it with always guessing the largest class.

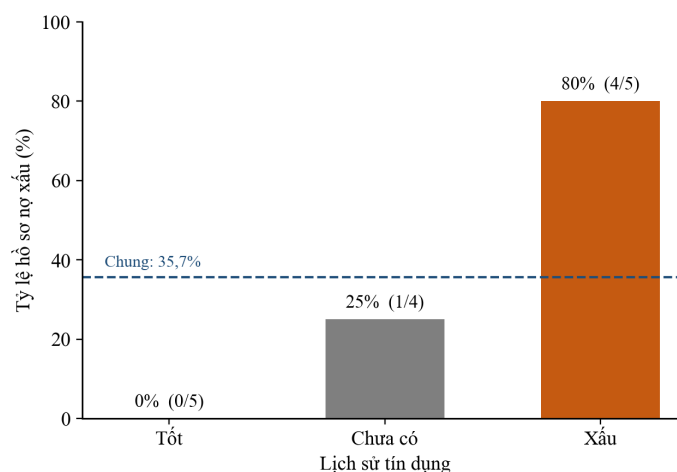
```
# (lịch sử tín dụng, nợ xấu: 1 = có) của 14 hồ sơ HS01 đến HS14
ho_so = [("Xấu", 1), ("Chưa có", 1), ("Xấu", 1), ("Xấu", 1),
         ("Xấu", 1), ("Xấu", 0), ("Tốt", 0), ("Tốt", 0),
         ("Tốt", 0), ("Tốt", 0), ("Tốt", 0), ("Chưa có", 0),
         ("Chưa có", 0), ("Chưa có", 0)]
nhom = {}
for ls, xau in ho_so:
    n, x = nhom.get(ls, (0, 0))
    nhom[ls] = (n + 1, x + xau)
for ls, (n, x) in nhom.items():
    print(f"{ls:8s} hồ sơ: {n}, nợ xấu: {x}, tỷ lệ {x / n:.0%}")
tong = sum(x for _, x in nhom.values())
print(f"Chung {tong}/{len(ho_so)} = {tong / len(ho_so):.1%}")
dung = sum(max(x, n - x) for n, x in nhom.values())
print(f"Đoán theo đa số của nhóm: {dung}/{len(ho_so)}")
mac_dinh = max(tong, len(ho_so) - tong)
print(f"Luôn đoán lớp đông nhất: {mac_dinh}/{len(ho_so)}")
```

Kết quả chạy / Output

```
Xấu      hồ sơ: 5, nợ xấu: 4, tỷ lệ 80%
Chưa có   hồ sơ: 4, nợ xấu: 1, tỷ lệ 25%
Tốt       hồ sơ: 5, nợ xấu: 0, tỷ lệ 0%
Chung     5/14 = 35.7%
Đoán theo đa số của nhóm: 12/14
Luôn đoán lớp đông nhất: 9/14
```

Nhóm lịch sử <i>History group</i>	Số hồ sơ <i>Applications</i>	Nợ xấu <i>Bad debt</i>	Tỷ lệ <i>Rate</i>	Lớp đoán (đa số) <i>Predicted class (majority)</i>	Đoán đúng <i>Correct</i>
Xấu <i>Bad</i>	5	4	80%	Nợ xấu <i>Bad debt</i>	4
Chưa có <i>None</i>	4	1	25%	Không nợ xấu <i>No bad debt</i>	3
Tốt <i>Good</i>	5	0	0%	Không nợ xấu <i>No bad debt</i>	5
Cộng <i>Total</i>	14	5	35,7%		12

- Đếm số hồ sơ và số hồ sơ nợ xấu của từng nhóm: Xấu 5 và 4, Chưa có 4 và 1, Tốt 5 và 0. Tổng 14 hồ sơ, 5 nợ xấu.
Count the applications and the bad-debt applications in each group: Bad 5 and 4, None 4 and 1, Good 5 and 0. In total 14 applications, 5 with bad debt.
- Tính tỷ lệ từng nhóm: $4/5 = 80\%$, $1/4 = 25\%$, $0/5 = 0\%$. Tỷ lệ chung $5/14 = 35,7\%$.
Compute each group's rate: $4/5 = 80\%$, $1/4 = 25\%$, $0/5 = 0\%$. The overall rate is $5/14 = 35.7\%$.
- Lập quy tắc: mỗi nhóm đoán theo lớp chiếm đa số. Nhóm Xấu đoán nợ xấu; hai nhóm còn lại đoán không nợ xấu.
Form the rule: each group is predicted as its majority class. The Bad group is predicted as bad debt; the other two groups as no bad debt.
- Số hồ sơ đoán đúng là $4 + 3 + 5 = 12$ trên 14, độ chính xác $85,7\%$. Cách luôn đoán "không nợ xấu" (9 hồ sơ) đúng $9/14 = 64,3\%$.
The number of correct predictions is $4 + 3 + 5 = 12$ out of 14, an accuracy of 85.7% . Always guessing "no bad debt" (9 applications) is right $9/14 = 64.3\%$ of the time.



Hình 1.1. Tỷ lệ hồ sơ nợ xấu theo lịch sử tín dụng trong 14 hồ sơ vay; đường nét đứt là tỷ lệ chung

Figure 1.1. Share of applications with bad debt by credit history among 14 loan applications; the dashed line is the overall share

Quy tắc trên hơn cách đoán mặc định 21,4 điểm phần trăm (85,7% so với 64,3%). Mẫu chỉ có 14 hồ sơ nên kết luận này mới là giả thuyết; Chương 7 trình bày cách kiểm trên dữ liệu chưa dùng để lập quy tắc.

The rule beats the default guess by 21.4 percentage points (85.7% against 64.3%). With only 14 applications this is still a hypothesis; Chapter 7 shows how to test it on data not used to build the rule.

PHẢI THUỘC / MEMORISE — Dữ liệu là giá trị thô. Thông tin là dữ liệu đã tổng hợp để trả lời một câu hỏi. Tri thức là mẫu hay quy luật dùng được cho dữ liệu mới và phải được kiểm trên dữ liệu chưa dùng khi lập nó.

Data is a raw value. Information is data summarised to answer a question. Knowledge is a pattern or rule usable on new data, and it has to be tested on data not used to build it.

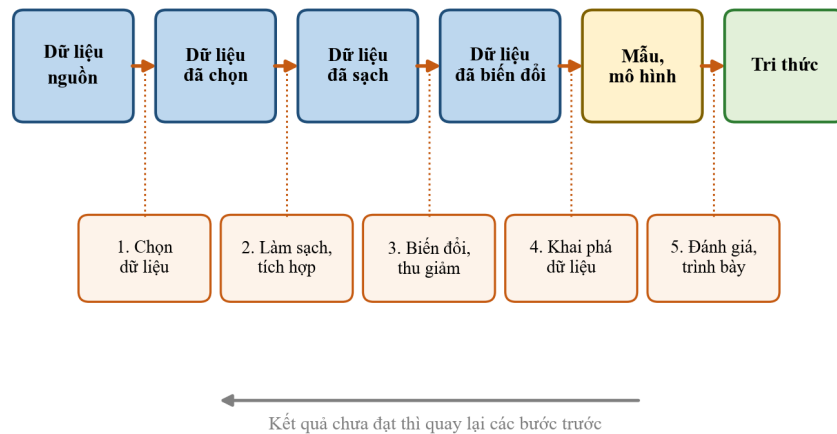
LỖI HAY GẶP / COMMON MISTAKE — Gọi mọi con số in ra là tri thức. "Có 12 khách" và "tổng dư nợ 415 triệu đồng" là thông tin; chỉ khi con số chuyển thành quy tắc dùng cho khách mới (ví dụ lịch sử Tốt thì ít rủi ro) và qua kiểm tra thì mới là tri thức.

Calling every printed number knowledge. "There are 12 customers" and "total debt 415 million dong" are information; only when a figure becomes a rule for new customers (for example a Good history means low risk) and passes a test is it knowledge.

1.2. Khai phá dữ liệu và quá trình KDD — 1.2. Data mining and the KDD process

Khai phá dữ liệu (data mining) là quá trình dùng thuật toán để tìm mẫu và mô hình có ích trong tập dữ liệu lớn, thay cho việc người phân tích tự nêu từng giả thuyết rồi kiểm từng cái. Tên gọi dễ gây hiểu nhầm: thứ được khai phá là tri thức nằm trong dữ liệu, giống như khai thác vàng từ quặng. Khám phá tri thức trong cơ sở dữ liệu (knowledge discovery in databases, KDD) là cả quá trình từ dữ liệu nguồn đến tri thức; khai phá dữ liệu là bước thứ tư trong năm bước. Nhiều tài liệu dùng hai tên như nhau, nhưng khi đề thi hỏi phân biệt thì khai phá dữ liệu chỉ là một bước của KDD.

Data mining is the process of using algorithms to find useful patterns and models in large data sets, instead of the analyst stating each hypothesis and testing them one by one. The name is misleading: what is mined is the knowledge lying in the data, as gold is mined from ore. Knowledge discovery in databases (KDD) is the whole process from source data to knowledge; data mining is the fourth of its five steps. Many texts use the two names interchangeably, but when an exam asks for the difference, data mining is only one step of KDD.



Hình 1.2. Quá trình KDD: sáu dạng dữ liệu nối nhau bằng năm bước, bước thứ tư là khai phá dữ liệu

Figure 1.2. The KDD process: six forms of data joined by five steps, the fourth being data mining

Bước Step	Việc làm What is done	Ví dụ ở Phú Sa Example at Phú Sa
1. Chọn dữ liệu <i>1. Select data</i>	Xác định bảng, cột, khoảng thời gian liên quan đến bài toán <i>Identify the tables, columns and period relevant to the problem</i>	Lấy thu nhập, nghề, hạng thẻ, lịch sử trả nợ; bỏ cột ghi chú tự do <i>Take income, occupation, card tier and repayment history; drop the free-text note</i>
2. Làm sạch, tích hợp <i>2. Clean, integrate</i>	Xử lý giá trị thiếu, nhiễu, trùng; ghép nhiều nguồn về một bảng <i>Handle missing values, noise and duplicates; merge several sources into one table</i>	Điền thu nhập thiếu của KH07; ghép bảng khách với bảng khoản vay theo mã khách <i>Fill the missing income of KH07; join the customer and loan tables on customer code</i>
3. Biến đổi, thu giảm <i>3. Transform, reduce</i>	Chuẩn hóa, rời rạc hóa, tạo thuộc tính mới, giảm số chiều <i>Normalise, discretise, derive new attributes, reduce dimensions</i>	Tính DTI (nợ trên thu nhập); chia thu nhập thành ba mức <i>Compute DTI (debt over income); split income into three levels</i>
4. Khai phá dữ liệu <i>4. Data mining</i>	Chạy thuật toán phân lớp, phân cụm, luật kết hợp để tìm mẫu <i>Run classification, clustering or association algorithms to find patterns</i>	Dựng mô hình đoán nợ xấu cho hồ sơ mới <i>Build a model that predicts bad debt for a new application</i>
5. Đánh giá, trình bày <i>5. Evaluate, present</i>	Loại mẫu tầm thường, đo độ chính xác, trình bày bằng bảng và hình để người dùng quyết định <i>Discard trivial patterns, measure accuracy, present with tables and figures for the user to decide</i>	Báo cáo cho khối tín dụng: ngưỡng điểm duyệt, tỷ lệ bỏ sót nợ xấu <i>Report to the credit division: the approval score cut-off and the share of missed bad debts</i>

Một mẫu tìm được chỉ có giá trị khi đạt bốn điều kiện: hợp lệ trên dữ liệu mới (valid), chưa biết trước (novel), có thể dùng vào việc cụ thể (potentially useful) và người dùng

hiểu được (understandable). Quy tắc "khách có khoản vay thì có dư nợ" hợp lệ nhưng tầm thường, nên bị loại ở bước 5.

A pattern found has value only if it meets four conditions: valid on new data, not known beforehand (novel), usable for a concrete task (potentially useful) and understandable to the user. The rule "a customer with a loan has outstanding debt" is valid but trivial, so it is discarded in step 5.

Khai phá dữ liệu khác truy vấn SQL (môn Cơ sở dữ liệu) ở loại câu hỏi. Truy vấn trả lời câu hỏi người dùng đã biết cách diễn đạt; khai phá tìm điều người dùng chưa nghĩ tới để hỏi. Cùng một bảng khách hàng, câu truy vấn sau đếm số khách và tổng dư nợ theo hạng thẻ:

Data mining differs from an SQL query (the Databases course) in the kind of question. A query answers a question the user can already phrase; mining finds what the user has not thought to ask. On the same customer table, the following query counts customers and total debt by card tier:

```
SELECT hang_the, COUNT(*) AS so_khach, SUM(du_no) AS tong_no
FROM khach_hang
GROUP BY hang_the;
```

Tiêu chí <i>Criterion</i>	Truy vấn SQL <i>SQL query</i>	Thống kê mô tả <i>Descriptive statistics</i>	Khai phá dữ liệu <i>Data mining</i>
Câu hỏi <i>Question</i>	Đã biết, viết thành điều kiện <i>Known, written as a condition</i>	Đã biết, hỏi về đặc trưng của tập dữ liệu <i>Known, about the features of the data set</i>	Chưa biết trước; thuật toán đề xuất mẫu <i>Not known beforehand; the algorithm proposes patterns</i>
Kết quả <i>Result</i>	Các dòng hay số tổng hợp <i>Rows or aggregate figures</i>	Trung bình, độ phân tán, biểu đồ <i>Means, spread, charts</i>	Quy tắc, mô hình, nhóm, điểm dự đoán <i>Rules, models, groups, predicted scores</i>
Ví dụ ở Phú Sa <i>Example at Phú Sa</i>	Đếm khách hạng Vàng <i>Count Gold-tier customers</i>	Thu nhập trung vị 15 triệu đồng <i>Median income 15 million dong</i>	Dự đoán hồ sơ mới có thành nợ xấu không <i>Predict whether a new application becomes bad debt</i>

Mã sau chạy lại cả năm bước trên bằng 12 khách ở thu nhỏ: chọn ba cột, làm sạch bằng cách điền thu nhập thiếu bằng trung vị, biến đổi thu nhập thành ba mức, rồi khai phá bằng cách đếm khách có nhà theo mức.

The following code runs the five steps again on the 12-customer table in miniature: select three columns, clean by filling the missing income with the median, transform income into three levels, then mine by counting house owners per level.

```
import statistics

# (mã, thu nhập triệu/tháng, có nhà); KH07 thiếu thu nhập
du_lieu = [("KH01", 9, 0), ("KH02", 12, 0), ("KH03", 15, 1),
           ("KH04", 22, 1), ("KH05", 8, 0), ("KH06", 18, 1),
           ("KH07", None, 1), ("KH08", 14, 0), ("KH09", 11, 0),
           ("KH10", 16, 1), ("KH11", 20, 1), ("KH12", 120, 1)]
# Làm sạch: điền ô thiếu bằng trung vị của các ô có giá trị
```

```
co_gia_tri = [t for _, t, _ in du_lieu if t is not None]
tv = statistics.median(co_gia_tri)
sach = [(m, tv if t is None else t, h) for m, t, h in du_lieu]

# Biến đổi: rời rạc hóa thu nhập thành ba mức
def muc(t):
    return "Thấp" if t < 12 else "Trung bình" if t < 20 else "Cao"

# Khai phá: đếm khách có nhà theo mức thu nhập
dem = {}
for m, t, h in sach:
    n, c = dem.get(muc(t), (0, 0))
    dem[muc(t)] = (n + 1, c + h)
print("Trung vị dùng để điền:", tv)
for k in ("Thấp", "Trung bình", "Cao"):
    n, c = dem[k]
    print(f"{k:10s} {n} khách, {c} có nhà ({c / n:.0%})")
```

Kết quả chạy / Output

```
Trung vị dùng để điền: 15
Thấp      3 khách, 0 có nhà (0%)
Trung bình 6 khách, 4 có nhà (67%)
Cao       3 khách, 3 có nhà (100%)
```

Trung vị của 11 thu nhập có giá trị là 15 triệu đồng, nên ô thiếu của KH07 được điền 15. Kết quả cho một mẫu có thể đọc ra: tỷ lệ có nhà tăng theo mức thu nhập, 0 trên 3, 4 trên 6 (67%) và 3 trên 3. Với 12 khách, đây là gợi ý cần kiểm, chưa phải kết luận. Chương 2 và 3 tính từng bước này kỹ hơn.

The median of the 11 observed incomes is 15 million dong, so the missing cell of KH07 is filled with 15. The run yields a readable pattern: the house-ownership rate rises with the income level, 0 of 3, 4 of 6 (67%) and 3 of 3. With 12 customers this is a lead to be tested, not a conclusion. Chapters 2 and 3 work through each of these steps in more detail.

PHẢI THUỘC / MEMORISE — KDD là cả quá trình năm bước; khai phá dữ liệu là bước thứ tư. Mẫu có giá trị khi hợp lệ, mới, hữu ích và hiểu được.

KDD is the whole five-step process; data mining is the fourth step. A pattern has value when it is valid, novel, useful and understandable.

LỖI HAY GẶP / COMMON MISTAKE — Viết "khai phá dữ liệu gồm làm sạch, chọn, biến đổi và tìm mẫu". Đó là mô tả KDD. Khai phá dữ liệu theo nghĩa hẹp chỉ là bước chạy thuật toán tìm mẫu; các bước còn lại là chuẩn bị và đánh giá.

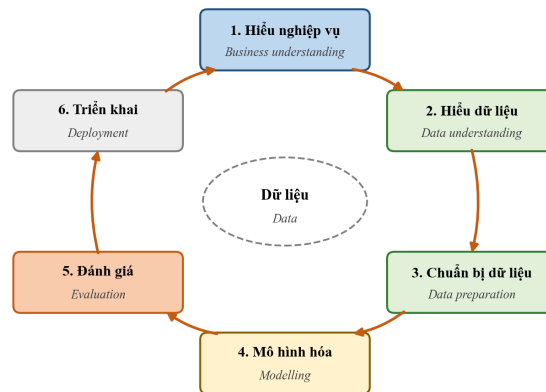
Writing "data mining consists of cleaning, selecting, transforming and finding patterns". That describes KDD. Data mining in the narrow sense is only the step that runs the pattern-finding algorithm; the other steps are preparation and evaluation.

1.3. CRISP-DM và dự án chấm điểm tín dụng của Phú Sa — 1.3. CRISP-DM and Phú Sa's credit scoring project

CRISP-DM (Cross-Industry Standard Process for Data Mining) là quy trình chuẩn cho dự án khai phá dữ liệu, gồm sáu pha xếp thành vòng. Khác với KDD tả luồng biến đổi của dữ liệu, CRISP-DM tả việc của nhóm dự án và nhấn mạnh nghiệp vụ ở đầu và cuối vòng.

Vòng lặp là chủ ý: gặp dữ liệu không đủ thì quay lại pha hiểu nghiệp vụ để đổi mục tiêu, mô hình chưa đạt thì quay lại pha chuẩn bị dữ liệu.

CRISP-DM (Cross-Industry Standard Process for Data Mining) is the standard process for a data mining project, with six phases arranged in a cycle. Where KDD describes how the data is transformed, CRISP-DM describes the work of the project team and puts the business at both the start and the end of the cycle. The loop is deliberate: when the data turns out to be insufficient the team returns to business understanding to change the goal, and when a model falls short it returns to data preparation.



Hình 1.3. Vòng CRISP-DM: sáu pha quanh dữ liệu, có thể quay lại pha trước

Figure 1.3. The CRISP-DM cycle: six phases around the data, with the option to return to an earlier phase

Áp vào Phú Sa: khối dữ liệu do giám đốc Phạm Thu Hà phụ trách xây mô hình chấm điểm tín dụng cho khoản vay tiền mặt trên ứng dụng "Phú Sa". Chuyên viên phân tích Đỗ Quang Huy dựng bảng dữ liệu và chạy mô hình; thực tập sinh Lê Ngọc Trâm kiểm đối chiếu nhãn nợ xấu với sổ thu hồi nợ. Bảng sau ghi việc của từng pha.

Applied to Phú Sa: the data division led by director Phạm Thu Hà builds a credit scoring model for cash loans on the "Phú Sa" mobile app. Analyst Đỗ Quang Huy builds the data table and runs the models; intern Lê Ngọc Trâm checks the bad-debt labels against the collections ledger. The table records the work of each phase.

Pha Phase	Việc chính Main tasks	Ở Phú Sa At Phú Sa	Chương g Chapter
1. Hiểu nghiệp vụ <i>1. Business understanding</i>	Nêu mục tiêu kinh doanh, tiêu chí thành công, ràng buộc <i>State the business goal, success criteria and constraints</i>	Giảm số khoản vay thành nợ xấu mà không từ chối nhiều khách tốt <i>Reduce loans turning into bad debt without rejecting many good customers</i>	1, 9

Pha <i>Phase</i>	Việc chính <i>Main tasks</i>	Ở Phú Sa <i>At Phú Sa</i>	Chương <i>Chapter</i>
2. Hiểu dữ liệu <i>2. Data understanding</i>	Thu thập, mô tả, soát chất lượng dữ liệu <i>Collect, describe and check the quality of the data</i>	Thống kê thu nhập, dư nợ; phát hiện thu nhập thiếu và ngoại lai <i>Summarise income and debt; find missing incomes and outliers</i>	2
3. Chuẩn bị dữ liệu <i>3. Data preparation</i>	Làm sạch, tích hợp, biến đổi, chọn thuộc tính, chia tập huấn luyện và kiểm tra <i>Clean, integrate, transform, select attributes, split into training and test sets</i>	Điền thu nhập thiếu, tính DTI, chuẩn hóa, chia tập hồ sơ <i>Fill missing income, compute DTI, normalise, split the applications</i>	3
4. Mô hình hóa <i>4. Modelling</i>	Chọn kỹ thuật, dựng và điều chỉnh mô hình <i>Choose techniques, build and tune models</i>	Cây quyết định, Naive Bayes, kNN, hồi quy logistic <i>Decision tree, Naive Bayes, kNN, logistic regression</i>	5, 6
5. Đánh giá <i>5. Evaluation</i>	Đo mô hình trên tập kiểm tra, đối chiếu với tiêu chí nghiệp vụ <i>Measure the model on the test set and compare with the business criteria</i>	Ma trận nhầm lẫn, ROC, chi phí bỏ sót nợ xấu <i>Confusion matrix, ROC, cost of missed bad debt</i>	7
6. Triển khai <i>6. Deployment</i>	Đưa mô hình vào hệ thống, giám sát, huấn luyện lại, tuân thủ pháp luật <i>Put the model into the system, monitor it, retrain it, comply with the law</i>	Chấm điểm hồ sơ trên ứng dụng, theo dõi độ ổn định điểm theo tháng <i>Score applications in the app and track score stability month by month</i>	7, 9

Chia dữ liệu thuộc pha 3 và phải làm trước khi dựng mô hình: mô hình học trên tập huấn luyện và chỉ được chấm điểm trên tập kiểm tra mà nó chưa thấy. Mã sau chia 14 hồ sơ vay thành tập kiểm tra 4 hồ sơ bằng bộ sinh số ngẫu nhiên có hạt giống 2026, nên lần chạy nào cũng ra cùng kết quả.

Splitting the data belongs to phase 3 and must be done before the model is built: the model learns on the training set and is scored only on a test set it has not seen. The following code splits the 14 loan applications into a test set of 4 using a random generator with seed 2026, so every run gives the same result.

```
import random

ma = [f"HS{i:02d}" for i in range(1, 15)]
nhan = dict(zip(ma, [1] * 5 + [0] * 9)) # 1 = nợ xấu
rng = random.Random(2026)
kiem_tra = sorted(rng.sample(ma, 4))
huan_luyen = [m for m in ma if m not in kiem_tra]
print("Kiểm tra:", kiem_tra)
print("Huấn luyện:", len(huan_luyen), "hồ sơ")
for ten, tap in (("huấn luyện", huan_luyen), ("kiểm tra", kiem_tra)):
    xau = sum(nhan[m] for m in tap)
    print(f"Nợ xấu trong tập {ten}: {xau}/{len(tap)}")
```

Kết quả chạy / Output

```
Kiểm tra: ['HS02', 'HS06', 'HS09', 'HS12']
Huấn luyện: 10 hồ sơ
Nợ xấu trong tập huấn luyện: 4/10
Nợ xấu trong tập kiểm tra: 1/4
```

Tập kiểm tra gồm HS02, HS06, HS09 và HS12; tập huấn luyện có 10 hồ sơ. Số hồ sơ nợ xấu là 4 trong tập huấn luyện và 1 trong tập kiểm tra, cộng lại đủ 5. Với tập nhỏ như vậy, phép chia ngẫu nhiên có thể làm lệch tỷ lệ lớp giữa hai tập; Chương 3 và 7 trình bày cách chia phân tầng.

The test set holds HS02, HS06, HS09 and HS12; the training set has 10 applications. Bad-debt applications number 4 in the training set and 1 in the test set, adding up to 5. With a set this small a random split can skew the class ratio between the two sets; Chapters 3 and 7 show stratified splitting.

PHẢI THUỘC / MEMORISE — Sáu pha CRISP-DM theo thứ tự: hiểu nghiệp vụ, hiểu dữ liệu, chuẩn bị dữ liệu, mô hình hóa, đánh giá, triển khai. Quá trình lặp, không đi một chiều. *The six CRISP-DM phases in order: business understanding, data understanding, data preparation, modelling, evaluation, deployment. The process iterates; it does not run in one direction.*

LỖI HAY GẶP / COMMON MISTAKE — Vào thẳng mô hình hóa mà chưa có tiêu chí thành công. Khi chưa nói "tốt" nghĩa là gì (giảm bao nhiêu nợ xấu, bỏ sót tối đa bao nhiêu), không thể quyết định mô hình nào đạt ở pha đánh giá. *Jumping into modelling without a success criterion. Until "good" has been defined (how much bad debt is to be cut, how many misses are tolerable), the evaluation phase cannot decide which model passes.*

1.4. Nhiệm vụ mô tả, nhiệm vụ dự đoán và các bài toán khai phá — 1.4.

Descriptive and predictive tasks and the mining problems

Nhiệm vụ khai phá chia làm hai nhóm. Nhiệm vụ mô tả (descriptive) tìm đặc trưng chung của dữ liệu hiện có: các nhóm khách giống nhau, các sản phẩm hay mua cùng nhau, điểm khác thường. Nhiệm vụ dự đoán (predictive) dùng dữ liệu đã biết kết quả để suy ra giá trị chưa biết của dữ liệu mới: khách này có trả nợ đúng hạn không, dư nợ tháng sau là bao nhiêu. Bảng sau xếp bảy bài toán thường gặp theo câu hỏi kinh doanh; thuật toán trong cột bốn được trình bày ở các chương ghi cột cuối.

Mining tasks fall into two groups. Descriptive tasks find general features of the existing data: groups of similar customers, products bought together, unusual points. Predictive tasks use data

with known outcomes to infer the unknown value for new data: whether this customer will repay on time, what next month's outstanding debt will be. The table sets out seven common problems by business question; the algorithms in column four are covered in the chapters named in the last column.

Câu hỏi kinh doanh <i>Business question</i>	Bài toán <i>Problem</i>	Thuật toán tiêu biểu <i>Typical algorithms</i>	Ví dụ ở Phú Sa <i>Example at Phú Sa</i>	Chương <i>Chapter</i>
Hồ sơ này có thành nợ xấu không? <i>Will this application become bad debt?</i>	Phân lớp <i>Classification</i>	Cây quyết định, Naive Bayes, kNN, hồi quy logistic <i>Decision tree, Naive Bayes, kNN, logistic regression</i>	Đoán nhãn Tốt hoặc Xấu cho hồ sơ vay mới <i>Predict the label Good or Bad for a new loan application</i>	5, 6
Dư nợ tháng sau là bao nhiêu? <i>What will next month's debt be?</i>	Hồi quy <i>Regression</i>	Hồi quy tuyến tính <i>Linear regression</i>	Dự báo dư nợ toàn công ty cho tháng 10/2026 <i>Forecast the company's outstanding debt for October 2026</i>	6
Khách nào giống khách nào? <i>Which customers resemble each other?</i>	Phân cụm <i>Clustering</i>	k-means, phân cấp, DBSCAN <i>k-means, hierarchical, DBSCAN</i>	Chia khách theo số giao dịch và chi tiêu <i>Group customers by transactions and spending</i>	8
Sản phẩm nào hay mua cùng nhau? <i>Which products are bought together?</i>	Luật kết hợp <i>Association rules</i>	Apriori, FP-growth <i>Apriori, FP-growth</i>	Khách trả góp xe máy đều mua kèm bảo hiểm khoản vay (3 trên 3 giao dịch) <i>Every motorbike instalment customer also buys loan insurance (3 of 3 transactions)</i>	4
Giao dịch nào khác thường? <i>Which transaction is unusual?</i>	Phát hiện bất thường <i>Anomaly detection</i>	z-score, IQR, khoảng cách, isolation forest <i>z-score, IQR, distance, isolation forest</i>	Giao dịch thẻ 4.200 nghìn đồng giữa các giao dịch vài trăm nghìn <i>A card transaction of 4,200 thousand dong among transactions of a few hundred thousand</i>	9

Câu hỏi kinh doanh <i>Business question</i>	Bài toán <i>Problem</i>	Thuật toán tiêu biểu <i>Typical algorithms</i>	Ví dụ ở Phú Sa <i>Example at Phú Sa</i>	Chương <i>Chapter</i>
Khách thường dùng sản phẩm theo thứ tự nào? <i>In what order do customers take products?</i>	Khai phá chuỗi <i>Sequence mining</i>	Mẫu tuần tự <i>Sequential patterns</i>	Vay tiền mặt, rồi thẻ, rồi bảo hiểm <i>Cash loan, then card, then insurance</i>	4
Khách nói gì về dịch vụ? <i>What do customers say about the service?</i>	Khai phá văn bản <i>Text mining</i>	TF-IDF, cosine, phân tích cảm xúc <i>TF-IDF, cosine, sentiment analysis</i>	Nhóm phản hồi về lãi suất và thủ tục <i>Group feedback about interest rates and procedures</i>	9

Hai bài toán đầu là dự đoán, năm bài toán sau là mô tả. Phân lớp và hồi quy giống nhau ở chỗ cần dữ liệu đã có đáp án; chúng khác nhau ở kiểu đáp án: phân lớp trả về một nhãn rời rạc, hồi quy trả về một số liên tục. Phân lớp và phân cụm thường bị nhầm vì kết quả của cả hai đều là các nhóm.

The first two problems are predictive and the other five are descriptive. Classification and regression are alike in needing data with known answers; they differ in the kind of answer: classification returns a discrete label, regression a continuous number. Classification and clustering are often confused because both produce groups.

Tiêu chí <i>Criterion</i>	Phân lớp <i>Classification</i>	Phân cụm <i>Clustering</i>
Nhãn có sẵn <i>Labels given</i>	Có, mỗi dòng huấn luyện có nhãn Tốt hoặc Xấu <i>Yes, each training row has the label Good or Bad</i>	Không, nhóm do thuật toán tạo ra <i>No, the groups are created by the algorithm</i>
Học theo kiểu <i>Learning type</i>	Có giám sát <i>Supervised</i>	Không giám sát <i>Unsupervised</i>
Đầu ra <i>Output</i>	Nhãn cho dòng mới; có thể đo đúng hay sai <i>A label for a new row; right or wrong can be measured</i>	Phân hoạch dữ liệu thành cụm; chất lượng đo gián tiếp <i>A partition of the data into clusters; quality measured indirectly</i>
Ví dụ ở Phú Sa <i>Example at Phú Sa</i>	Hồ sơ vay mới là Tốt hay Xấu <i>Is a new loan application Good or Bad</i>	Chia khách theo mức giao dịch và chi tiêu rồi đặt tên nhóm <i>Group customers by activity and spending, then name the groups</i>

PHẢI THUỘC / MEMORISE — Phân lớp và hồi quy cần nhãn (có giám sát), một trả nhãn rời rạc, một trả số. Phân cụm, luật kết hợp, phát hiện bất thường không cần nhãn.
Classification and regression need labels (supervised), one returning a discrete label and the other a number. Clustering, association rules and anomaly detection need no labels.

LỖI HAY GẶP / COMMON MISTAKE — Dùng phân lớp khi chưa có nhãn. Nếu Phú Sa chưa ghi nhận khoản vay nào là nợ xấu hay không, không có dữ liệu huấn luyện để phân lớp; việc đầu tiên là định nghĩa và gắn nhãn, hoặc dùng phân cụm để khám phá.
Using classification when no labels exist. If Phú Sa has not recorded which loans became bad debt, there is no training data for classification; the first job is to define and attach the labels, or to use clustering to explore.

1.5. Học có giám sát, không giám sát và bán giám sát — 1.5. Supervised, unsupervised and semi-supervised learning

Cách thuật toán học phụ thuộc vào việc dữ liệu có nhãn hay không. Học có giám sát (supervised learning) dùng các dòng đã biết đáp án, gọi là nhãn, để học quy tắc từ thuộc tính ra nhãn; phân lớp và hồi quy thuộc nhóm này. Học không giám sát (unsupervised learning) chỉ có thuộc tính, không có nhãn, và tìm cấu trúc trong dữ liệu: phân cụm, luật kết hợp, phát hiện bất thường. Học bán giám sát (semi-supervised learning) dùng một phần nhỏ dòng có nhãn cùng nhiều dòng chưa có nhãn.

How an algorithm learns depends on whether the data is labelled. Supervised learning uses rows with a known answer, called the label, to learn a rule from attributes to label; classification and regression belong here. Unsupervised learning has attributes only, no labels, and looks for structure in the data: clustering, association rules, anomaly detection. Semi-supervised learning uses a small share of labelled rows together with many unlabelled ones.

Phú Sa gặp cả ba tình huống. Hồ sơ vay đã đáo hạn có nhãn Tốt hoặc Xấu, đủ để học có giám sát. Dữ liệu giao dịch thẻ không có nhãn gian lận sẵn nên phải dùng phương pháp không giám sát để tìm giao dịch khác thường. Khoản vay vừa giải ngân chưa biết sẽ thành nợ xấu hay không, nên là dòng chưa có nhãn; khi nhãn khan hiếm và việc gắn nhãn tốn công, học bán giám sát tận dụng cả hai loại dòng.

Phú Sa meets all three situations. Loan applications that have matured carry the label Good or Bad, enough for supervised learning. Card transaction data has no ready-made fraud label, so unsupervised methods are needed to find unusual transactions. A loan just disbursed is not yet known to become bad debt, so it is an unlabelled row; when labels are scarce and costly to attach, semi-supervised learning makes use of both kinds of rows.

Tiêu chí Criterion	Có giám sát Supervised	Không giám sát Unsupervised	Bán giám sát Semi-supervised
Dữ liệu Data	Mọi dòng có nhãn Every row labelled	Không có nhãn No labels	Ít dòng có nhãn, nhiều dòng không Few rows labelled, many not

Tiêu chí <i>Criterion</i>	Có giám sát <i>Supervised</i>	Không giám sát <i>Unsupervised</i>	Bán giám sát <i>Semi-supervised</i>
Mục tiêu <i>Goal</i>	Đoán nhãn cho dòng mới <i>Predict the label of a new row</i>	Tìm nhóm, quy luật, điểm lạ <i>Find groups, rules, odd points</i>	Đoán nhãn, dùng cấu trúc của dòng chưa nhãn <i>Predict labels, using the structure of unlabelled rows</i>
Thuật toán <i>Algorithms</i>	Cây quyết định, kNN, hồi quy <i>Decision tree, kNN, regression</i>	k-means, Apriori, DBSCAN <i>k-means, Apriori, DBSCAN</i>	Tự gán nhãn, lan truyền nhãn <i>Self-training, label propagation</i>
Ví dụ ở Phú Sa <i>Example at Phú Sa</i>	Đoán nợ xấu từ hồ sơ đã đáo hạn <i>Predict bad debt from matured applications</i>	Chia khách theo hành vi dùng ứng dụng <i>Group customers by app behaviour</i>	Khoản vay chưa đủ kỳ hạn để biết nhãn <i>Loans too young for the label to be known</i>

Ví dụ nhỏ nhất của học có giám sát là quy tắc một ngưỡng. Mười hồ sơ của bộ dữ liệu ở Chương 6 có thu nhập, tỷ lệ nợ trên thu nhập (DTI, tính bằng phần trăm) và nhãn Tốt hay Xấu. Quy tắc: DTI lớn hơn 40 thì đoán Xấu, ngược lại đoán Tốt. Mã sau áp quy tắc cho cả mười hồ sơ, đếm bốn ô của bảng đúng sai và so với cách luôn đoán lớp đông nhất.

The smallest example of supervised learning is a one-threshold rule. The ten applications of the Chapter 6 data set carry income, debt-to-income ratio (DTI, in percent) and the label Good or Bad. The rule: DTI above 40 means predict Bad, otherwise predict Good. The following code applies the rule to all ten, counts the four cells of the right-and-wrong table and compares with always guessing the largest class.

Bài mẫu 2. Chấm quy tắc ngưỡng DTI trên 40 — *Worked example 2. Scoring the rule DTI above 40*

```
# (mã, thu nhập triệu/tháng, DTI %, nhãn thật)
td = [("TD01", 9, 52, "Xấu"), ("TD02", 11, 46, "Xấu"),
      ("TD03", 14, 40, "Tốt"), ("TD04", 16, 30, "Tốt"),
      ("TD05", 20, 25, "Tốt"), ("TD06", 26, 18, "Tốt"),
      ("TD07", 13, 44, "Xấu"), ("TD08", 18, 38, "Xấu"),
      ("TD09", 22, 33, "Tốt"), ("TD10", 10, 35, "Tốt")]

def quy_tac(dti):
    return "Xấu" if dti > 40 else "Tốt"

bang = {"TP": 0, "FN": 0, "FP": 0, "TN": 0}
dung = 0
for ma, thu, dti, that in td:
    doan = quy_tac(dti)
    dung += doan == that
    if that == "Xấu":
        bang["TP" if doan == "Xấu" else "FN"] += 1
    else:
        bang["FP" if doan == "Xấu" else "TN"] += 1
    if doan != that:
        print("Sai:", ma, "DTI", dti, "đoán", doan, "thật", that)
print(bang)
```

```
print(f"Đúng {dung}/{len(td)} = {dung / len(td):.0%}")
nhan = [x[3] for x in td]
print("Luôn đoán Tốt:", nhan.count("Tốt") / len(td))
```

Kết quả chạy / Output

```
Sai: TD08 DTI 38 đoán Tốt thật Xấu
{'TP': 3, 'FN': 1, 'FP': 0, 'TN': 6}
Đúng 9/10 = 90%
Luôn đoán Tốt: 0.6
```

Mã Code	DTI (%)	Nhân thật Actual	Quy tắc đoán Rule predicts	Kết quả Outcome
TD01	52	Xấu <i>Bad</i>	Xấu <i>Bad</i>	Đúng <i>Right</i>
TD02	46	Xấu <i>Bad</i>	Xấu <i>Bad</i>	Đúng <i>Right</i>
TD03	40	Tốt <i>Good</i>	Tốt <i>Good</i>	Đúng <i>Right</i>
TD04	30	Tốt <i>Good</i>	Tốt <i>Good</i>	Đúng <i>Right</i>
TD05	25	Tốt <i>Good</i>	Tốt <i>Good</i>	Đúng <i>Right</i>
TD06	18	Tốt <i>Good</i>	Tốt <i>Good</i>	Đúng <i>Right</i>
TD07	44	Xấu <i>Bad</i>	Xấu <i>Bad</i>	Đúng <i>Right</i>
TD08	38	Xấu <i>Bad</i>	Tốt <i>Good</i>	Sai <i>Wrong</i>
TD09	33	Tốt <i>Good</i>	Tốt <i>Good</i>	Đúng <i>Right</i>
TD10	35	Tốt <i>Good</i>	Tốt <i>Good</i>	Đúng <i>Right</i>

- Với từng hồ sơ so DTI với 40: TD01 (52), TD02 (46) và TD07 (44) lớn hơn 40 nên bị đoán Xấu; bảy hồ sơ còn lại bị đoán Tốt (TD03 có DTI đúng bằng 40, không lớn hơn).

For each application compare DTI with 40: TD01 (52), TD02 (46) and TD07 (44) exceed 40 and are predicted Bad; the other seven are predicted Good (TD03 has a DTI of exactly 40, not above it).

- So với nhân thật: ba hồ sơ bị đoán Xấu đều thật sự Xấu; sáu hồ sơ Tốt đều được đoán Tốt; riêng TD08 (DTI 38) thật sự Xấu nhưng bị đoán Tốt.

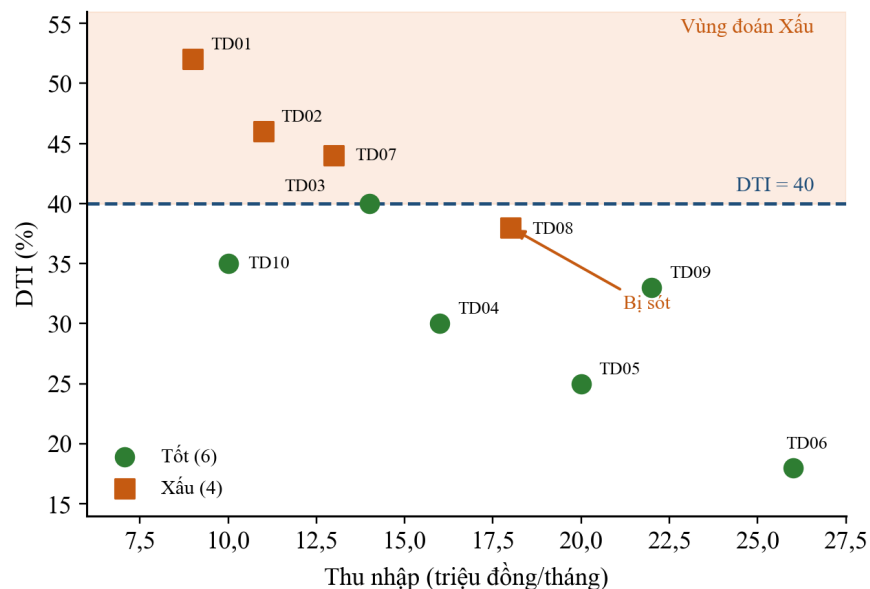
Compared with the actual labels: the three predicted Bad are all truly Bad; all six Good applications are predicted Good; only TD08 (DTI 38) is truly Bad but predicted Good.

3. Gọi Xấu là lớp cần tìm: TP = 3 (đoán Xấu, thật Xấu), FN = 1 (đoán Tốt, thật Xấu), FP = 0 (đoán Xấu, thật Tốt), TN = 6.

Take Bad as the class to find: TP = 3 (predicted Bad, truly Bad), FN = 1 (predicted Good, truly Bad), FP = 0 (predicted Bad, truly Good), TN = 6.

4. Độ chính xác = $(TP + TN) / 10 = (3 + 6) / 10 = 90\%$. Cách luôn đoán Tốt đúng 6/10 = 60% (đúng 0,6 trong kết quả chạy). Quy tắc hơn 30 điểm phần trăm.

Accuracy = $(TP + TN) / 10 = (3 + 6) / 10 = 90\%$. Always guessing Good is right 6/10 = 60% (0.6 in the run output). The rule is ahead by 30 percentage points.



Hình 1.4. Mười hồ sơ theo thu nhập và DTI; vùng tô là nơi quy tắc đoán Xấu, TD08 là hồ sơ Xấu bị sót

Figure 1.4. Ten applications by income and DTI; the shaded area is where the rule predicts Bad, and TD08 is the Bad application the rule misses

Độ chính xác 90% đo trên chính mười hồ sơ dùng để chọn ngưỡng, nên lạc quan. Ngưỡng 36 cũng cho 90% trên mười hồ sơ này (bỏ sót 0 hồ sơ Xấu nhưng đoán nhầm TD03), tức dữ liệu chưa đủ để phân biệt hai ngưỡng. Chương 6 và 7 trình bày cách chọn ngưỡng và đo trên tập kiểm tra tách riêng.

The 90% accuracy is measured on the same ten applications used to choose the threshold, so it is optimistic. A threshold of 36 also gives 90% on these ten (missing no Bad application but misjudging TD03), so the data cannot separate the two thresholds. Chapters 6 and 7 show how to choose a threshold and measure on a separate test set.

PHẢI THUỘC / MEMORISE — Có giám sát: có nhãn, học ánh xạ từ thuộc tính ra nhãn. Không giám sát: không nhãn, tìm cấu trúc. Bán giám sát: ít nhãn, nhiều dòng chưa nhãn.
Supervised: labels exist, learn the mapping from attributes to label. Unsupervised: no labels, find structure. Semi-supervised: few labels, many unlabelled rows.

LỖI HAY GẶP / COMMON MISTAKE — Báo độ chính xác đo trên chính dữ liệu đã dùng để chọn ngưỡng hay dựng mô hình. Con số 90% ở trên chỉ nói quy tắc khớp với mười

hồ sơ; khả năng đoán hồ sơ mới phải đo trên tập kiểm tra.

Reporting accuracy measured on the same data used to choose the threshold or build the model. The 90% above says only that the rule fits the ten applications; the ability to predict new applications must be measured on a test set.

1.6. Các loại dữ liệu được khai phá — 1.6. Types of data that are mined

Dữ liệu có cấu trúc (structured) nằm trong bảng có cột và kiểu cố định; dữ liệu bán cấu trúc (semi-structured) như JSON, XML có thể nhưng độ dài và thứ tự thay đổi; dữ liệu phi cấu trúc (unstructured) như văn bản tự do, ảnh, âm thanh không có khung sẵn. Theo cách tổ chức, năm loại sau gặp nhiều trong khai phá dữ liệu tài chính.

Structured data sits in tables with fixed columns and types; semi-structured data such as JSON and XML has tags but variable length and order; unstructured data such as free text, images and audio has no ready frame. By how it is organised, the following five types are common in financial data mining.

Loại Type	Cách tổ chức Organisation	Ví dụ ở Phú Sa Example at Phú Sa	Bài toán hay dùng Typical problem	Chương Chapter
Bảng quan hệ <i>Relational table</i>	Mỗi dòng là một đối tượng, mỗi cột là một thuộc tính <i>Each row an object, each column an attribute</i>	Bảng 12 khách hàng; 14 hồ sơ vay <i>The 12-customer table; the 14 loan applications</i>	Phân lớp, phân cụm, hồi quy <i>Classification, clustering, regression</i>	2 đến 8
Giao dịch <i>Transaction</i>	Mỗi bản ghi là một tập mục đi cùng nhau <i>Each record is a set of items occurring together</i>	10 giao dịch, mỗi giao dịch là các sản phẩm khách dùng <i>10 transactions, each the products a customer uses</i>	Luật kết hợp <i>Association rules</i>	4
Chuỗi thời gian <i>Time series</i>	Giá trị đo lặp theo thứ tự thời gian <i>Values measured repeatedly in time order</i>	Dư nợ cuối tháng từ 10/2025 đến 9/2026 <i>Month-end debt from October 2025 to September 2026</i>	Dự báo, phát hiện bất thường <i>Forecasting, anomaly detection</i>	6, 9
Văn bản <i>Text</i>	Chuỗi từ, cần tách từ và vector hóa <i>Sequences of words, requiring tokenisation and vectorisation</i>	Bốn phản hồi của khách về lãi suất và thủ tục <i>Four customer comments on interest rates and procedures</i>	Phân nhóm, tìm tương tự, cảm xúc <i>Grouping, similarity, sentiment</i>	9

Loại Type	Cách tổ chức Organisation	Ví dụ ở Phú Sa Example at Phú Sa	Bài toán hay dùng Typical problem	Chương Chapter
Đồ thị Graph	Đỉnh là thực thể, cạnh là quan hệ <i>Nodes are entities, edges are relationships</i>	Hai khách cùng đăng nhập từ một thiết bị là hai đỉnh nối nhau <i>Two customers logging in from one device are two linked nodes</i>	Phát hiện gian lận, lan truyền rủi ro <i>Fraud detection, risk propagation</i>	9

Giao dịch và chuỗi thời gian có mã chạy ngay ở đây. Dữ liệu giao dịch ghi mỗi lần khách mua gói sản phẩm; mã sau đếm tần suất từng sản phẩm trong 10 giao dịch (BH: bảo hiểm khoản vay; TD: thẻ tín dụng; VM: vay tiền mặt; DM: trả góp điện máy; XM: trả góp xe máy).

Transactions and time series have code that runs here. Transaction data records each time a customer takes a bundle of products; the following code counts how often each product appears in 10 transactions (BH: loan insurance; TD: credit card; VM: cash loan; DM: electronics instalment; XM: motorbike instalment).

```
from collections import Counter

giao_dich = [
    {"TD", "VM", "BH"}, {"TD", "BH"}, {"VM", "DM", "BH"},
    {"TD", "VM", "BH"}, {"XM", "BH"}, {"TD", "DM"},
    {"VM", "XM", "BH"}, {"TD", "VM"}, {"DM", "XM", "BH"},
    {"TD", "VM", "BH"},
]
n = len(giao_dich)
dem = Counter(sp for g in giao_dich for sp in g)
for sp, c in sorted(dem.items(), key=lambda kv: (-kv[1], kv[0])):
    print(f"{sp}: {c}/{n} = {c / n:.0%}")
tb = sum(len(g) for g in giao_dich) / n
print("Số sản phẩm mỗi giao dịch:", tb)
```

Kết quả chạy / Output

```
BH: 8/10 = 80%
TD: 6/10 = 60%
VM: 6/10 = 60%
DM: 3/10 = 30%
XM: 3/10 = 30%
Số sản phẩm mỗi giao dịch: 2.6
```

Bảo hiểm khoản vay có mặt trong 8 trong 10 giao dịch (80%), thẻ và vay tiền mặt mỗi loại 60%, hai loại trả góp mỗi loại 30%. Mỗi giao dịch có trung bình 2,6 sản phẩm. Dữ liệu này không có cột thuộc tính, chỉ có tập mục, nên cần thuật toán riêng (Chương 4).

Loan insurance appears in 8 of 10 transactions (80%), the card and the cash loan 60% each, the two instalment products 30% each. Each transaction holds 2.6 products on average. This data has no attribute columns, only item sets, so it needs its own algorithms (Chapter 4).

Chuỗi thời gian có thứ tự là thông tin: đổi chỗ hai tháng là làm mất xu hướng. Mã sau đọc 12 giá trị dư nợ cuối tháng (tỷ đồng) và tính mức thay đổi từng tháng.

A time series carries information in its order: swapping two months destroys the trend. The following code reads 12 month-end debt values (billion dong) and computes the change in each month.

```
thang = ["10/2025", "11/2025", "12/2025", "01/2026", "02/2026",
        "03/2026", "04/2026", "05/2026", "06/2026", "07/2026",
        "08/2026", "09/2026"]
du_no = [820.5, 838.2, 851.7, 879.3, 884.1, 912.6,
        925.8, 948.2, 951.7, 987.4, 1003.9, 1030.5] # tỷ đồng
chenh = [round(b - a, 1) for a, b in zip(du_no, du_no[1:])]
print("Chênh lệch từng tháng (tỷ đồng):")
print(" ", chenh[:6])
print(" ", chenh[6:])
print("Tăng cả kỳ:", round(du_no[-1] - du_no[0], 1))
print(f"Tăng tương đối: {du_no[-1] / du_no[0] - 1:.1%}")
print("Trung bình mỗi tháng:", round(sum(chenh) / len(chenh), 2))
lon = max(chenh)
print("Tháng tăng nhiều nhất:", thang[chenh.index(lon) + 1], lon)
```

Kết quả chạy / Output

```
Chênh lệch từng tháng (tỷ đồng):
[17.7, 13.5, 27.6, 4.8, 28.5, 13.2]
[22.4, 3.5, 35.7, 16.5, 26.6]
Tăng cả kỳ: 210.0
Tăng tương đối: 25.6%
Trung bình mỗi tháng: 19.09
Tháng tăng nhiều nhất: 07/2026 35.7
```

Dư nợ tăng từ 820,5 lên 1.030,5 tỷ đồng, thêm 210,0 tỷ đồng (25,6%) trong 11 bước tháng, trung bình 19,09 tỷ đồng mỗi tháng. Tháng tăng nhiều nhất là 07/2026 với 35,7 tỷ đồng. Chuỗi luôn đi lên nhưng không đều, nên dự báo cần mô hình xu hướng (Chương 6) và phép làm trơn (Chương 9).

Debt rises from 820.5 to 1,030.5 billion dong, an increase of 210.0 billion dong (25.6%) over 11 monthly steps, 19.09 billion dong a month on average. The largest monthly rise is July 2026 at 35.7 billion dong. The series always rises but unevenly, so forecasting needs a trend model (Chapter 6) and smoothing (Chapter 9).

Dữ liệu lớn (big data) là dữ liệu mà khối lượng, tốc độ hoặc độ đa dạng vượt khả năng của công cụ thông thường. Mô hình 3V gồm khối lượng (volume), tốc độ (velocity), đa dạng (variety); mô hình 5V thêm độ tin cậy (veracity) và giá trị (value).

Big data is data whose volume, speed or variety exceeds what ordinary tools can handle. The 3V model covers volume, velocity and variety; the 5V model adds veracity and value.

V	Ý nghĩa Meaning	Biểu hiện ở Phú Sa How it shows at Phú Sa
Khối lượng Volume	Số bản ghi và dung lượng lớn <i>A large number of records and a large size</i>	Mỗi lần mở ứng dụng, xem khoản vay, bấm thanh toán đều thành một bản ghi sự kiện <i>Every app opening, loan view and payment tap becomes an event record</i>
Tốc độ Velocity	Dữ liệu đến liên tục, cần xử lý nhanh <i>Data arrives continuously and must be processed fast</i>	Chấm điểm hồ sơ ngay khi khách nộp trên ứng dụng <i>Scoring an application the moment the customer submits it in the app</i>

V	Ý nghĩa <i>Meaning</i>	Biểu hiện ở Phú Sa <i>How it shows at Phú Sa</i>
Đa dạng <i>Variety</i>	Nhiều dạng: bảng, văn bản, ảnh, chuỗi sự kiện <i>Many forms: tables, text, images, event streams</i>	Bảng khoản vay, phản hồi tự do, ảnh giấy tờ định danh <i>Loan tables, free-text feedback, images of identity papers</i>
Tin cậy <i>Veracity</i>	Mức đúng, đủ, nhất quán của dữ liệu <i>How accurate, complete and consistent the data is</i>	Thu nhập tự khai thiếu hoặc sai, mã khách trùng <i>Self-declared income missing or wrong, duplicated customer codes</i>
Giá trị <i>Value</i>	Dữ liệu phải đem lại lợi ích kinh doanh đo được <i>The data must bring a measurable business benefit</i>	Tiền nợ xấu tránh được nhờ chấm điểm tốt hơn <i>Bad debt avoided by better scoring</i>

PHẢI THUỘC / MEMORISE — Ba V gốc: khối lượng, tốc độ, đa dạng; hai V thêm: độ tin cậy, giá trị. Dữ liệu giao dịch là tập mục, chuỗi thời gian có thứ tự là thông tin, văn bản cần tách từ.

The three original Vs: volume, velocity, variety; the two added Vs: veracity, value. Transaction data is a set of items, a time series carries information in its order, text needs tokenising.

LỖI HAY GẶP / COMMON MISTAKE — Xáo trộn hay chia ngẫu nhiên chuỗi thời gian như dữ liệu bảng. Tháng tương lai lọt vào tập huấn luyện làm kết quả đẹp giả; với chuỗi thời gian, tập kiểm tra phải là các tháng sau tập huấn luyện.

Shuffling or randomly splitting a time series as if it were a table. Future months leaking into the training set produce a falsely good result; for a time series the test set must be the months after the training set.

1.7. Kho dữ liệu, OLAP và sự khác biệt với khai phá dữ liệu — 1.7. Data warehouse, OLAP and how they differ from data mining

Hệ thống giao dịch (OLTP, online transaction processing) ghi từng nghiệp vụ như giải ngân, thu nợ; dữ liệu thay đổi liên tục và tối ưu cho ghi nhanh. Kho dữ liệu (data warehouse) là nơi gom dữ liệu lịch sử từ nhiều hệ thống OLTP để phân tích, có bốn đặc điểm: hướng chủ đề (subject-oriented), tích hợp (integrated), theo thời gian (time-variant) và ổn định (non-volatile, ghi thêm chứ không sửa). Việc đưa dữ liệu vào kho gọi là ETL (extract, transform, load: trích, biến đổi, nạp), tương ứng bước làm sạch và tích hợp của KDD.

A transaction system (OLTP, online transaction processing) records each operation such as a disbursement or a repayment; its data changes continuously and is tuned for fast writes. A data warehouse gathers historical data from several OLTP systems for analysis and has four features: subject-oriented, integrated, time-variant and non-volatile (appended to, not edited). Loading data into the warehouse is called ETL (extract, transform, load), matching the cleaning and integration step of KDD.

OLAP (online analytical processing) là cách hỏi đáp nhanh trên kho dữ liệu theo các chiều. Chiều (dimension) là góc nhìn như nghề, hạng thẻ, tháng; số đo (measure) là con số được tổng hợp như số khách, dư nợ. Năm phép thao tác thường gặp: roll-up (gộp lên mức tổng quát hơn, ví dụ từ tháng lên quý), drill-down (xuống mức chi tiết hơn), slice (cắt một giá trị của một chiều), dice (cắt theo nhiều chiều), pivot (xoay trục để đổi cách nhìn).

OLAP (online analytical processing) is fast question answering on a warehouse along dimensions. A dimension is a viewpoint such as occupation, card tier or month; a measure is a summarised figure such as customer count or outstanding debt. Five common operations: roll-up (aggregate to a more general level, for example from month to quarter), drill-down (go to a finer level), slice (fix one value of one dimension), dice (select on several dimensions), pivot (rotate the axes to change the view).

Mã sau dựng khối hai chiều nghề và hạng thẻ từ 12 khách, in số khách từng ô rồi thực hiện roll-up theo nghề để có dư nợ theo hạng thẻ (triệu đồng).

The following code builds a two-dimensional cube of occupation and card tier from the 12 customers, prints the customer count in each cell, then rolls up over occupation to get debt by card tier (million dong).

```
from collections import Counter

khach = [ # (nghề, hạng thẻ, dư nợ triệu đồng)
    ("Công nhân", "Chuẩn", 12), ("Nhân viên văn phòng", "Chuẩn", 18),
    ("Giáo viên", "Vàng", 20), ("Kinh doanh tự do", "Vàng", 45),
    ("Công nhân", "Chuẩn", 10), ("Nhân viên văn phòng", "Vàng", 30),
    ("Kinh doanh tự do", "Vàng", 40), ("Giáo viên", "Chuẩn", 16),
    ("Nhân viên văn phòng", "Chuẩn", 14), ("Công nhân", "Vàng", 25),
    ("Giáo viên", "Bạch kim", 35), ("Kinh doanh tự do", "Bạch kim", 150),
]
hang = ["Chuẩn", "Vàng", "Bạch kim"]
nghe = sorted({k[0] for k in khach})
dem = Counter({k[0], k[1] for k in khach})
print(f'Nghề:22s' + " ".join(f'{h:>10s}' for h in hang))
for ng in nghe:
    print(f'{ng:22s}' + " ".join(f'{dem[(ng, h)]:>10d}' for h in hang))
# Roll-up: gộp mọi nghề, cộng dư nợ theo hạng thẻ
no = Counter()
for _, h, d in khach:
    no[h] += d
print("Dư nợ theo hạng:", {h: no[h] for h in hang})
print("Tổng dư nợ:", sum(no.values()), "triệu đồng")
```

Kết quả chạy / Output

Nghề	Chuẩn	Vàng	Bạch kim
Công nhân	2	1	0
Giáo viên	1	1	1
Kinh doanh tự do	0	2	1
Nhân viên văn phòng	2	1	0
Dư nợ theo hạng:	{ 'Chuẩn': 70, 'Vàng': 160, 'Bạch kim': 185 }		
Tổng dư nợ:	415 triệu đồng		

Hạng Bạch kim có 2 trong 12 khách (16,7%) nhưng giữ 185 trong 415 triệu đồng dư nợ (44,6%). Đó là câu trả lời OLAP: nó cho biết dư nợ tập trung ở đâu. Nó không nói khách Bạch kim nào sẽ trả chậm; câu hỏi đó cần mô hình dự đoán ở Chương 5 và 6.

The Platinum tier has 2 of 12 customers (16.7%) but holds 185 of 415 million dong of debt (44.6%). That is an OLAP answer: it says where the debt is concentrated. It does not say which Platinum customer will pay late; that question needs the predictive models of Chapters 5 and 6.

Tiêu chí <i>Criterion</i>	OLAP <i>OLAP</i>	Khai phá dữ liệu <i>Data mining</i>
Hướng tiếp cận <i>Approach</i>	Người dùng chọn chiều và phép toán, hệ thống tổng hợp <i>The user chooses dimensions and operations; the system aggregates</i>	Thuật toán tự tìm mẫu từ dữ liệu <i>The algorithm finds patterns from the data itself</i>
Câu hỏi <i>Question</i>	Bao nhiêu, ở đâu, khi nào <i>How many, where, when</i>	Vì sao, ai sẽ, nhóm nào, cái gì đi cùng cái gì <i>Why, who will, which group, what goes with what</i>
Kết quả <i>Result</i>	Bảng tổng hợp, khối dữ liệu <i>Summary tables, cubes</i>	Quy tắc, mô hình, cụm, điểm dự đoán <i>Rules, models, clusters, predicted scores</i>
Ví dụ ở Phú Sa <i>Example at Phú Sa</i>	Dư nợ theo hạng thẻ và quý <i>Debt by card tier and quarter</i>	Khách nào có xác suất nợ xấu cao <i>Which customers have a high probability of bad debt</i>

PHẢI THUỘC / MEMORISE — OLAP tổng hợp theo chiều do người dùng chọn và trả lời "bao nhiêu, ở đâu". Khai phá dữ liệu tự tìm mẫu và trả lời "vì sao, ai sẽ". Kho dữ liệu là nguồn thuận lợi cho cả hai nhưng không bắt buộc phải có.

OLAP aggregates along dimensions the user chooses and answers "how many, where". Data mining finds patterns itself and answers "why, who will". A warehouse is a convenient source for both but is not mandatory.

LỖI HAY GẶP / COMMON MISTAKE — Cho rằng có kho dữ liệu và biểu đồ OLAP là đã khai phá dữ liệu. Biểu đồ roll-up chỉ tổng hợp; thiếu bước tìm mẫu và kiểm mẫu trên dữ liệu mới thì chưa phải khai phá.

Believing that having a warehouse and OLAP charts amounts to data mining. A roll-up chart only aggregates; without a step that finds patterns and tests them on new data it is not mining.

1.8. Ứng dụng trong tài chính và công cụ khai phá dữ liệu — 1.8. Applications in finance and data mining tools

Sáu ứng dụng sau là các bài toán Phú Sa gặp hằng ngày; mỗi bài toán quy về một hay nhiều bài toán ở mục 1.4.

The following six applications are problems Phú Sa meets every day; each reduces to one or more of the problems in section 1.4.

Ứng dụng <i>Application</i>	Câu hỏi <i>Question</i>	Bài toán <i>Problem</i>	Dữ liệu dùng <i>Data used</i>	Chương <i>Chapter</i>
Chấm điểm tín dụng <i>Credit scoring</i>	Khách này có trả nợ đúng hạn không? <i>Will this customer repay on time?</i>	Phân lớp, xác suất <i>Classification, probability</i>	Thu nhập, DTI, lịch sử trả nợ <i>Income, DTI, repayment history</i>	5, 6, 7
Chống gian lận <i>Fraud prevention</i>	Giao dịch hay hồ sơ nào khác thường? <i>Which transaction or application is unusual?</i>	Phát hiện bất thường <i>Anomaly detection</i>	Giá trị, thời điểm, thiết bị giao dịch <i>Value, time and device of transactions</i>	9
Bán chéo <i>Cross-selling</i>	Khách này nên được mời sản phẩm nào? <i>Which product should this customer be offered?</i>	Luật kết hợp <i>Association rules</i>	Giỏ sản phẩm của từng khách <i>Each customer's product basket</i>	4
Phân khúc khách hàng <i>Customer segmentation</i>	Có mấy nhóm khách, mỗi nhóm cần chính sách nào? <i>How many customer groups are there, and what policy does each need?</i>	Phân cụm, RFM <i>Clustering, RFM</i>	Số giao dịch, chi tiêu, thời gian từ lần cuối <i>Number of transactions, spending, time since the last one</i>	8
Dự báo dòng tiền <i>Cash-flow forecasting</i>	Dư nợ và tiền thu tháng tới ra sao? <i>What will next month's debt and collections be?</i>	Hồi quy, chuỗi thời gian <i>Regression, time series</i>	Dư nợ, tiền thu theo tháng <i>Monthly debt and collections</i>	6, 9
Phòng chống rửa tiền <i>Anti-money laundering</i>	Giao dịch nào cần chuyên viên tuân thủ kiểm tra? <i>Which transaction needs a compliance officer's review?</i>	Bất thường, luật cảnh báo <i>Anomalies, alert rules</i>	Chuỗi giao dịch, quan hệ giữa các tài khoản <i>Transaction sequences, links between accounts</i>	9

Ứng dụng cuối có người trong vòng quyết định: mô hình chỉ đề xuất cảnh báo, chuyên viên tuân thủ quyết định có báo cáo hay không. Ngưỡng cảnh báo đặt thấp thì bắt nhiều giao dịch lạ nhưng tạo nhiều cảnh báo sai, đặt cao thì ngược lại; đánh đổi này nói tiếp ở Chương 7 và 9.

The last application keeps a person in the loop: the model only proposes alerts and a compliance officer decides whether to report. A low alert threshold catches many unusual transactions but raises many false alerts, a high one does the reverse; this trade-off is taken up in Chapters 7 and 9.

Về công cụ, sách này viết thuật toán thuần Python bằng thư viện chuẩn để người học tính tay và kiểm lại từng bước. Trong dự án thật, nhóm dùng thư viện có sẵn; mỗi công cụ ở bảng sau chỉ được mô tả cách dùng, không so sánh hơn kém.

As for tools, this book writes the algorithms in plain Python using the standard library so that learners can compute by hand and check each step. In a real project the team uses ready-made libraries; each tool in the table is described by how it is used, not ranked.

Công cụ <i>Tool</i>	Loại <i>Kind</i>	Cách dùng <i>How it is used</i>	Thành phần nhắc đến trong sách <i>Components named in this book</i>
Python, scikit-learn, pandas	Ngôn ngữ lập trình và thư viện <i>Programming language and libraries</i>	Viết mã dựng quy trình từ làm sạch đến mô hình; pandas giữ bảng dữ liệu, scikit-learn cung cấp thuật toán <i>Code builds the pipeline from cleaning to model; pandas holds the data table, scikit-learn supplies algorithms</i>	DecisionTreeClassifier, GaussianNB, KMeans
R	Ngôn ngữ thống kê <i>Statistical language</i>	Phân tích thống kê và dựng mô hình bằng gói và hàm viết sẵn <i>Statistical analysis and modelling with ready-made packages and functions</i>	Các gói thống kê và phân cụm <i>Statistics and clustering packages</i>
Weka	Phần mềm có giao diện đồ họa, viết bằng Java <i>Graphical software written in Java</i>	Nạp tệp dữ liệu, chọn thuật toán trong menu, đọc kết quả mà không cần viết mã <i>Load a data file, pick an algorithm from a menu, read the results without writing code</i>	Cây quyết định, Naive Bayes, kNN, k-means, Apriori <i>Decision tree, Naive Bayes, kNN, k-means, Apriori</i>
Orange	Phần mềm lập trình trực quan <i>Visual programming software</i>	Kéo và nối các khối thao tác thành quy trình, có thể mở rộng bằng Python <i>Drag and connect operation blocks into a workflow, extensible with Python</i>	Phân lớp, phân cụm, biểu đồ khám phá <i>Classification, clustering, exploratory charts</i>

PHẢI THUỘC / MEMORISE — Mỗi ứng dụng tài chính quy về bài toán khai phá: chấm điểm là phân lớp, chống gian lận và rửa tiền là phát hiện bất thường, bán chéo là luật kết hợp, phân khúc là phân cụm, dòng tiền là hồi quy và chuỗi thời gian.

Each financial application reduces to a mining problem: scoring is classification, fraud and money-laundering control is anomaly detection, cross-selling is association rules, segmentation is clustering, cash flow is regression and time series.

LỖI HAY GẶP / COMMON MISTAKE — Chọn công cụ trước rồi mới tìm bài toán. Công cụ chỉ thực thi thuật toán; bài toán và tiêu chí đánh giá phải xác định từ pha hiểu nghiệp vụ

(mục 1.3).

Choosing the tool first and the problem afterwards. A tool only executes algorithms; the problem and the evaluation criterion must come from the business understanding phase (section 1.3).

1.9. Quy trình dự án, vai trò nhân sự và thách thức — 1.9. Project process, team roles and challenges

Một dự án khai phá dữ liệu ở công ty tài chính chạy theo vòng CRISP-DM nhưng được quản lý như mọi dự án khác: có phạm vi, mốc bàn giao, người chịu trách nhiệm từng pha. Lặp lại thường quay về pha 3 khi dữ liệu chưa đủ sạch, và vòng đầu thường chỉ cần một mô hình đơn giản làm mốc so sánh. Sáu vai trò sau tham gia dự án chấm điểm tín dụng của Phú Sa.

A data mining project at a finance company runs on the CRISP-DM cycle but is managed like any other project: it has a scope, delivery milestones and an owner for each phase. Iterations often return to phase 3 when the data is not yet clean enough, and the first cycle usually needs only a simple model as a baseline. The following six roles take part in Phú Sa's credit scoring project.

Vai trò <i>Role</i>	Việc chính <i>Main work</i>	Ở Phú Sa <i>At Phú Sa</i>
Chủ nghiệp vụ <i>Business owner</i>	Đặt mục tiêu, tiêu chí thành công, nhận kết quả vào quy trình duyệt vay <i>Sets the goal and success criteria, takes the result into the lending process</i>	Lãnh đạo khối tín dụng <i>Head of the credit division</i>
Lãnh đạo khối dữ liệu <i>Head of data</i>	Phân việc, giữ tiến độ, báo cáo chủ nghiệp vụ <i>Assigns work, keeps the schedule, reports to the business owner</i>	Giám đốc Phạm Thu Hà <i>Director Phạm Thu Hà</i>
Kỹ sư dữ liệu <i>Data engineer</i>	Xây luồng ETL, kho dữ liệu, bảo đảm dữ liệu cập nhật <i>Builds the ETL flow and warehouse, keeps data up to date</i>	Nhóm hạ tầng dữ liệu <i>The data infrastructure team</i>
Nhà phân tích dữ liệu <i>Data analyst</i>	Khám phá dữ liệu, chuẩn bị bảng, dựng và đánh giá mô hình <i>Explores the data, prepares the table, builds and evaluates models</i>	Chuyên viên Đỗ Quang Huy; thực tập sinh Lê Ngọc Trâm hỗ trợ kiểm nhãn <i>Analyst Đỗ Quang Huy; intern Lê Ngọc Trâm helps check labels</i>
Chuyên viên tuân thủ <i>Compliance officer</i>	Soát cơ sở pháp lý xử lý dữ liệu cá nhân, nội dung xin đồng ý <i>Reviews the legal basis for processing personal data and the consent wording</i>	Bộ phận pháp chế, tuân thủ <i>The legal and compliance unit</i>

Vai trò <i>Role</i>	Việc chính <i>Main work</i>	Ở Phú Sa <i>At Phú Sa</i>
Kiểm định mô hình <i>Model validator</i>	Kiểm độc lập mô hình trước khi đưa vào dùng, theo dõi sau triển khai <i>Independently checks the model before use and monitors it after deployment</i>	Bộ phận quản trị rủi ro <i>The risk management unit</i>

Năm thách thức sau lặp lại ở hầu hết dự án. Mỗi thách thức có biểu hiện cụ thể ở Phú Sa và một cách xử lý.

The following five challenges recur in most projects. Each has a concrete form at Phú Sa and a way of handling it.

Thách thức <i>Challenge</i>	Biểu hiện ở Phú Sa <i>How it shows at Phú Sa</i>	Cách xử lý <i>How to handle it</i>	Chương <i>Chapter</i>
Chất lượng dữ liệu <i>Data quality</i>	Thu nhập của KH07 thiếu; thu nhập 120 triệu đồng của KH12 cách xa nhóm còn lại <i>KH07's income is missing; KH12's income of 120 million dong lies far from the rest</i>	Soát, điền hoặc loại; kiểm ngoại lai theo quy tắc rõ ràng <i>Check, fill or drop; test outliers by an explicit rule</i>	2, 3
Quyền riêng tư <i>Privacy</i>	Hồ sơ vay chứa thông tin tài chính, tín dụng của khách <i>Loan files hold customers' financial and credit information</i>	Chỉ lấy cột cần dùng, dùng mã khách thay tên, giới hạn người truy cập <i>Take only the columns needed, use customer codes instead of names, limit access</i>	9
Giải thích được <i>Explainability</i>	Khách bị từ chối hỏi lý do; nhân viên cần nói được mô hình dựa vào đâu <i>A rejected customer asks why; staff must say what the model relied on</i>	Ưu tiên mô hình đọc được như cây quyết định, hồi quy; ghi lại các yếu tố ảnh hưởng <i>Prefer readable models such as trees and regression; record the factors that drive the score</i>	5, 6, 9
Quá khớp <i>Overfitting</i>	Mô hình đúng 99% trên hồ sơ cũ nhưng sai nhiều trên hồ sơ mới <i>A model 99% right on old applications but often wrong on new ones</i>	Chia tập huấn luyện và kiểm tra, kiểm chéo <i>Split into training and test sets, cross-validate</i>	7
Thiên lệch <i>Bias</i>	Nhãn nợ xấu chỉ có ở khách đã được duyệt, khách bị từ chối không có nhãn <i>Bad-debt labels exist only for approved customers; rejected ones have none</i>	Nhận diện nhóm thiếu dữ liệu, kiểm tác động theo nhóm khách <i>Identify the groups with missing data and test the effect by customer group</i>	9

PHẢI THUỘC / MEMORISE — Năm thách thức: chất lượng dữ liệu, quyền riêng tư, giải thích được, quá khớp, thiên lệch. Dự án chạy theo vòng, chủ nghiệp vụ đặt tiêu chí và người kiểm định độc lập chấm mô hình.

Five challenges: data quality, privacy, explainability, overfitting, bias. The project runs in a cycle; the business owner sets the criteria and an independent validator scores the model.

LỖI HAY GẶP / COMMON MISTAKE — Chỉ giao dự án cho người phân tích, bỏ chủ nghiệp vụ và pháp chế khỏi cuộc họp đầu. Mô hình đẹp về kỹ thuật nhưng không trả lời đúng câu hỏi kinh doanh, hoặc dùng cột dữ liệu mà công ty không được phép dùng, phải làm lại từ đầu.

Giving the project to the analyst alone and leaving the business owner and legal staff out of the first meeting. A model that is technically good but does not answer the business question, or uses a data column the company may not use, has to be redone from the start.

1.10. Khung pháp lý tóm lược — 1.10. A short legal framework

Khai phá dữ liệu khách hàng là xử lý dữ liệu, và với công ty tài chính còn là xử lý dữ liệu cá nhân và thông tin khách hàng. Bảng sau tóm tắt bốn nhóm quy định nền tảng, mỗi dòng ghi Điều đã đối chiếu với văn bản. Chi tiết, kèm Luật Trí tuệ nhân tạo, để Chương 9.

Mining customer data is data processing, and for a finance company also the processing of personal data and customer information. The table summarises four groups of basic rules, each row naming the Article checked against the text. Details, together with the Law on Artificial Intelligence, are left to Chapter 9.

Văn bản <i>Instrument</i>	Điều <i>Article</i>	Nội dung <i>Content</i>	Liên hệ với dự án <i>Link to the project</i>
Luật Dữ liệu 60/2024/QH15, hiệu lực 01/7/2025 (bản hợp nhất 42/VBHN-VPQH ngày 16/3/2026) <i>Law on Data 60/2024/QH15, effective 01/7/2025 (consolidated text 42/VBHN-VPQH of 16/3/2026)</i>	Điều 3 khoản 1, 8, 9 <i>Article 3, clauses 1, 8, 9</i>	Dữ liệu là dữ liệu số về sự vật, hiện tượng, sự kiện; xử lý dữ liệu gồm tiếp nhận, chuyển đổi, tổ chức dữ liệu và các hoạt động khác; cơ sở dữ liệu là tập hợp dữ liệu được sắp xếp để truy cập, khai thác <i>Data means digital data on things, phenomena and events; data processing covers receiving, converting and organising data and other activities; a database is a set of data arranged for access and exploitation</i>	Bảng khách hàng, kho dữ liệu và các bước làm sạch, biến đổi đều là xử lý dữ liệu <i>The customer table, the warehouse and the cleaning and transformation steps are all data processing</i>
Luật Bảo vệ dữ liệu cá nhân 91/2025/QH15,	Điều 2 khoản 1, 3, 11; Điều 3	Dữ liệu cá nhân gồm dữ liệu cơ bản và nhạy cảm, sau khi khử nhận dạng không còn là dữ liệu cá nhân;	Chỉ dùng dữ liệu khách cho mục đích đã nêu; mã khách còn bảng đối

Văn bản <i>Instrument</i>	Điều <i>Article</i>	Nội dung <i>Content</i>	Liên hệ với dự án <i>Link to the project</i>
hiệu lực 01/01/2026 <i>Law on Personal Data Protection 91/2025/QH15, effective 01/01/2026</i>	khoản 2 <i>Article 2, clauses 1, 3, 11; Article 3, clause 2</i>	dữ liệu nhạy cảm thuộc danh mục do Chính phủ ban hành; chỉ thu thập, xử lý đúng phạm vi, mục đích cụ thể, rõ ràng <i>Personal data comprises basic and sensitive data and stops being personal data once de-identified; sensitive data is in a list issued by the Government; collection and processing must stay within a specific and clear scope and purpose</i>	chỉ cho người thật chưa phải là khử nhận dạng <i>Use customer data only for the stated purpose; a customer code that still has a lookup table to a real person is not de-identification</i>
Nghị định 356/2025/NĐ-CP (31/12/2025, hiệu lực 01/01/2026; Nghị định 13/2023/NĐ-CP hết hiệu lực) <i>Decree 356/2025/ND-CP (31/12/2025, effective 01/01/2026; Decree 13/2023/ND-CP ceased to be in force)</i>	Điều 4 khoản 1 điểm k; Điều 8 khoản 2 điểm a; Điều 42 khoản 2 <i>Article 4, clause 1(k); Article 8, clause 2(a); Article 42, clause 2</i>	Thông tin tài chính, tín dụng và lịch sử giao dịch của khách tại tổ chức tín dụng là dữ liệu cá nhân nhạy cảm; khi xin đồng ý, tổ chức tài chính, ngân hàng phải nêu rõ mục đích, gồm chấm điểm, xếp hạng tín dụng nếu có <i>Financial and credit information and transaction history of customers at credit institutions is sensitive personal data; when seeking consent, financial and banking organisations must state the purposes, including credit scoring and rating where applicable</i>	Dữ liệu khoản vay của Phú Sa là dữ liệu nhạy cảm; nội dung xin đồng ý phải nói rõ có chấm điểm tín dụng <i>Phú Sa's loan data is sensitive data; the consent wording must say that credit scoring takes place</i>
Luật Các tổ chức tín dụng, bản hợp nhất 143/2026/VBH N-LQ-VPQH ngày 23/9/2026 <i>Law on Credit Institutions, consolidated text 143/2026/VBHN-LQ-VPQH of 23/9/2026</i>	Điều 4 khoản 41; Điều 13 khoản 1, 2, 3 <i>Article 4, clause 41; Article 13, clauses 1, 2, 3</i>	Công ty tài chính là loại hình tổ chức tín dụng phi ngân hàng. Người quản lý, nhân viên không được tiết lộ thông tin khách hàng; tổ chức tín dụng phải bảo đảm bí mật thông tin khách hàng và không cung cấp cho tổ chức, cá nhân khác, trừ yêu cầu của cơ quan nhà nước có thẩm quyền theo luật hoặc được khách hàng chấp thuận <i>A finance company is a type of non-bank credit institution. Managers and staff may not disclose customer information; a credit institution must keep customer information confidential and may not provide it to other organisations or individuals,</i>	Gửi dữ liệu khách cho nhà thầu hay công cụ bên ngoài công ty cần có căn cứ này; nhân viên không được tiết lộ thông tin khách <i>Sending customer data to a contractor or a tool outside the company needs this basis; staff may not disclose customer information</i>

Văn bản <i>Instrument</i>	Điều <i>Article</i>	Nội dung <i>Content</i>	Liên hệ với dự án <i>Link to the project</i>
		<i>except at the request of a competent state body under law or with the customer's consent</i>	

Hai điểm cần nhớ khi làm bài. Thứ nhất, Nghị định 356/2025 thay Nghị định 13/2023, nên không viện dẫn văn bản cũ như đang hiệu lực. Thứ hai, bản hợp nhất 143/2026 đã gộp Luật 23/2026/QH16 nhưng phần do Luật này sửa chỉ có hiệu lực từ 01/12/2026; Điều 13 nêu ở bảng là nội dung của bản hợp nhất hiện hành.

Two points to remember in an exam. First, Decree 356/2025 replaces Decree 13/2023, so the old text is not cited as in force. Second, the consolidated text 143/2026 has merged Law 23/2026/QH16 but the parts amended by that Law take effect only on 01/12/2026; Article 13 in the table is the content of the current consolidated text.

PHẢI THUỘC / MEMORISE — Thông tin tài chính, tín dụng của khách tại tổ chức tín dụng là dữ liệu cá nhân nhạy cảm (Nghị định 356/2025 Điều 4 khoản 1 điểm k). Tổ chức tín dụng phải giữ bí mật thông tin khách hàng (Luật Các tổ chức tín dụng Điều 13).
Financial and credit information of customers at credit institutions is sensitive personal data (Decree 356/2025, Article 4, clause 1(k)). A credit institution must keep customer information confidential (Law on Credit Institutions, Article 13).

LỖI HAY GẶP / COMMON MISTAKE — Gọi mọi bảng có mã khách là dữ liệu đã ẩn danh. Theo Luật Bảo vệ dữ liệu cá nhân Điều 2 khoản 11, khử nhận dạng là làm cho dữ liệu không thể xác định hoặc không thể giúp xác định một con người cụ thể; mã khách vẫn nói được về người thật qua bảng đối chiếu thì chưa đạt.
Calling every table with customer codes anonymised data. Under the Law on Personal Data Protection, Article 2, clause 11, de-identification makes data unable to identify or help identify a specific person; a customer code that can still be linked back to a real person through a lookup table does not qualify.

1.11. Bảng thuật ngữ — 1.11. Glossary

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Dữ liệu	data	Giá trị ghi lại về đối tượng hay sự kiện, chưa gắn câu hỏi <i>A recorded value about an object or event, not yet tied to a question</i>
Thông tin	information	Dữ liệu đã tổng hợp, đặt trong ngữ cảnh để trả lời một câu hỏi <i>Data summarised and placed in context to answer a question</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Tri thức	knowledge	Mẫu hay quy luật khái quát được cho dữ liệu mới <i>A pattern or rule that generalises to new data</i>
Khai phá dữ liệu	data mining	Dùng thuật toán tìm mẫu và mô hình có ích trong dữ liệu lớn <i>Using algorithms to find useful patterns and models in large data</i>
Khám phá tri thức trong cơ sở dữ liệu	knowledge discovery in databases (KDD)	Quá trình năm bước từ dữ liệu nguồn đến tri thức <i>The five-step process from source data to knowledge</i>
CRISP-DM	CRISP-DM	Quy trình chuẩn sáu pha cho dự án khai phá dữ liệu <i>The standard six-phase process for a data mining project</i>
Mô hình	model	Biểu diễn toán học học từ dữ liệu, dùng để dự đoán hoặc mô tả <i>A mathematical representation learned from data, used to predict or describe</i>
Thuộc tính	attribute	Một cột mô tả đối tượng, ví dụ thu nhập, nghề <i>A column describing an object, for example income or occupation</i>
Nhãn	label	Đáp án đã biết của một dòng, ví dụ Tốt hoặc Xấu <i>The known answer for a row, for example Good or Bad</i>
Phân lớp	classification	Dự đoán nhãn rời rạc cho dòng mới từ dòng đã có nhãn <i>Predicting a discrete label for a new row from labelled rows</i>
Hồi quy	regression	Dự đoán một giá trị số liên tục <i>Predicting a continuous numeric value</i>
Phân cụm	clustering	Chia dữ liệu thành nhóm giống nhau khi chưa có nhãn <i>Dividing data into similar groups when no labels exist</i>
Luật kết hợp	association rule	Quy tắc nói mục này hay xuất hiện cùng mục kia <i>A rule saying one item tends to occur with another</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Phát hiện bất thường	anomaly detection	Tìm điểm dữ liệu khác hẳn phần còn lại <i>Finding data points that differ sharply from the rest</i>
Học có giám sát	supervised learning	Học từ dòng có nhãn để đoán nhãn cho dòng mới <i>Learning from labelled rows to predict labels for new rows</i>
Học không giám sát	unsupervised learning	Tìm cấu trúc trong dữ liệu không có nhãn <i>Finding structure in data without labels</i>
Học bán giám sát	semi-supervised learning	Học từ ít dòng có nhãn cùng nhiều dòng chưa nhãn <i>Learning from few labelled rows plus many unlabelled ones</i>
Tập huấn luyện	training set	Phần dữ liệu để mô hình học <i>The part of the data the model learns from</i>
Tập kiểm tra	test set	Phần dữ liệu mô hình chưa thấy, dùng để chấm điểm <i>The part of the data the model has not seen, used for scoring</i>
Kho dữ liệu	data warehouse	Kho gom dữ liệu lịch sử từ nhiều hệ thống để phân tích <i>A store gathering historical data from several systems for analysis</i>
OLAP	online analytical processing (OLAP)	Tổng hợp nhanh theo nhiều chiều trên kho dữ liệu <i>Fast multidimensional aggregation on a warehouse</i>
ETL	extract, transform, load (ETL)	Trích, biến đổi, nạp dữ liệu vào kho <i>Extracting, transforming and loading data into the warehouse</i>
Dữ liệu lớn	big data	Dữ liệu vượt khả năng công cụ thường về khối lượng, tốc độ hay đa dạng <i>Data beyond ordinary tools in volume, velocity or variety</i>
Khử nhận dạng	de-identification	Thay đổi hoặc xóa thông tin để dữ liệu không còn xác định được con người cụ thể <i>Changing or removing information so data can no longer identify a specific person</i>

Thuật ngữ <i>Term (Vietnamese)</i>	Tiếng Anh <i>English</i>	Giải thích ngắn <i>Short definition</i>
Dữ liệu cá nhân nhạy cảm	sensitive personal data	Dữ liệu cá nhân gắn với quyền riêng tư, thuộc danh mục do Chính phủ ban hành <i>Personal data tied to privacy, in a list issued by the Government</i>

1.12. Bài tập vận dụng có lời giải — *Applied exercises with solutions*

Bài 1. Quy tắc theo kỳ hạn vay — *Exercise 1. A rule based on loan term*

Công ty Tài chính TNHH Phú Sa ghi lại 8 khoản trả góp điện máy (dữ kiện đề bài): C01 đến C04 kỳ hạn 6 tháng, trong đó chỉ C03 trễ hạn; C05 đến C08 kỳ hạn 12 tháng, trong đó C05, C07 và C08 trễ hạn. (a) Tính tỷ lệ trễ hạn của từng kỳ hạn và của cả 8 khoản. (b) Lập quy tắc dùng một thuộc tính là kỳ hạn, mỗi kỳ hạn đoán theo lớp chiếm đa số. (c) Tính độ chính xác của quy tắc và so với cách luôn đoán lớp đông nhất.

Phú Sa Finance Company Limited records 8 electronics instalment loans (problem data): C01 to C04 have a 6-month term and only C03 is late; C05 to C08 have a 12-month term and C05, C07 and C08 are late. (a) Compute the late rate of each term and of all 8 loans. (b) Form a one-attribute rule on the term, predicting each term by its majority class. (c) Compute the rule's accuracy and compare it with always guessing the largest class.

- Kỳ hạn 6 tháng có 4 khoản, 1 khoản trễ: $1/4 = 25\%$. Kỳ hạn 12 tháng có 4 khoản, 3 khoản trễ: $3/4 = 75\%$. Cả 8 khoản có $1 + 3 = 4$ khoản trễ: $4/8 = 50\%$.
The 6-month term has 4 loans, 1 late: $1/4 = 25\%$. The 12-month term has 4 loans, 3 late: $3/4 = 75\%$. All 8 loans have $1 + 3 = 4$ late: $4/8 = 50\%$.
- Quy tắc: kỳ hạn 6 tháng đoán không trễ (lớp đa số, 3 trên 4); kỳ hạn 12 tháng đoán trễ (lớp đa số, 3 trên 4).
Rule: the 6-month term is predicted not late (majority class, 3 of 4); the 12-month term is predicted late (majority class, 3 of 4).
- Số khoản đoán đúng: $3 + 3 = 6$ trên 8, độ chính xác $6/8 = 75\%$.
Correct predictions: $3 + 3 = 6$ out of 8, an accuracy of $6/8 = 75\%$.
- Lớp đông nhất: 4 khoản trễ và 4 khoản không trễ bằng nhau, nên luôn đoán một lớp đúng $4/8 = 50\%$. Quy tắc theo kỳ hạn hơn 25 điểm phần trăm; với 8 khoản, kết luận này cần kiểm trên khoản vay mới.
The largest class: 4 late and 4 not late are tied, so always guessing one class is right $4/8 = 50\%$. The term-based rule is ahead by 25 percentage points; with 8 loans the conclusion needs testing on new loans.

Bài 2. Làm sạch và rời rạc hóa thu nhập — *Exercise 2. Cleaning and discretising income*

Bảng 7 khách có thu nhập (triệu đồng mỗi tháng, dữ kiện đề bài) là 10, 14, thiếu, 8, 20, 12, 16. (a) Điền ô thiếu bằng trung vị của các ô có giá trị. (b) Rời rạc hóa thành ba mức: dưới 12 là Thấp, từ 12 đến dưới 18 là Trung bình, từ 18 trở lên là Cao. (c) Đếm số khách mỗi mức và nói bước này thuộc bước nào của KDD.

A table of 7 customers has incomes (million dong per month, problem data) of 10, 14, missing, 8, 20, 12, 16. (a) Fill the missing cell with the median of the observed cells. (b) Discretise into three levels: under 12 is Low, from 12 up to under 18 is Medium, 18 and above is High. (c) Count the customers in each level and say which KDD step each operation belongs to.

1. Các ô có giá trị sắp tăng dần: 8, 10, 12, 14, 16, 20. Có 6 giá trị (số chẵn), trung vị là trung bình hai giá trị giữa: $(12 + 14) / 2 = 13$. Ô thiếu được điền 13.
The observed cells in ascending order: 8, 10, 12, 14, 16, 20. There are 6 values (an even count), so the median is the mean of the two middle values: $(12 + 14) / 2 = 13$. The missing cell is filled with 13.
2. Dãy đã điền: 10, 14, 13, 8, 20, 12, 16. Gán mức: 10 là Thấp; 14 là Trung bình; 13 là Trung bình; 8 là Thấp; 20 là Cao; 12 là Trung bình; 16 là Trung bình.
The filled series: 10, 14, 13, 8, 20, 12, 16. Assign levels: 10 is Low; 14 is Medium; 13 is Medium; 8 is Low; 20 is High; 12 is Medium; 16 is Medium.
3. Đếm: Thấp 2 khách, Trung bình 4 khách, Cao 1 khách, tổng 7. Mức Trung bình chiếm $4/7 = 57,1\%$.
Count: Low 2 customers, Medium 4, High 1, total 7. The Medium level holds $4/7 = 57.1\%$.
4. Điền giá trị thiếu thuộc bước 2 (làm sạch, tích hợp); rời rạc hóa thuộc bước 3 (biến đổi, thu giảm); đếm theo mức là việc của bước 4 (khai phá) ở dạng đơn giản nhất.
Filling the missing value belongs to step 2 (clean, integrate); discretisation to step 3 (transform, reduce); counting by level is step 4 (mining) in its simplest form.

Bài 3. Dung lượng dữ liệu sự kiện ứng dụng — Exercise 3. Storage size of app event data

Mỗi ngày có 40.000 khách mở ứng dụng "Phú Sa"; mỗi khách sinh 5 bản ghi sự kiện; mỗi bản ghi dài 200 byte (dữ kiện đề bài, 1 MB = 1.000.000 byte). (a) Tính số bản ghi và dung lượng mỗi ngày. (b) Tính dung lượng sau 30 ngày và sau 365 ngày. (c) Chỉ ra chữ V nào của dữ liệu lớn thể hiện rõ nhất và loại dữ liệu theo cách tổ chức.

Each day 40,000 customers open the "Phú Sa" app; each customer generates 5 event records; each record is 200 bytes long (problem data, 1 MB = 1,000,000 bytes). (a) Compute the number of records and the storage per day. (b) Compute the storage after 30 days and after 365 days. (c) Name the V of big data that shows most clearly and the type of data by organisation.

1. Số bản ghi mỗi ngày: $40.000 * 5 = 200.000$ bản ghi.
Records per day: $40,000 * 5 = 200,000$ records.
2. Dung lượng mỗi ngày: $200.000 * 200 = 40.000.000$ byte = 40 MB.
Storage per day: $200,000 * 200 = 40,000,000$ bytes = 40 MB.
3. Sau 30 ngày: $40 * 30 = 1.200$ MB = 1,2 GB. Sau 365 ngày: $40 * 365 = 14.600$ MB = 14,6 GB (1 GB = 1.000 MB).
After 30 days: $40 * 30 = 1,200$ MB = 1.2 GB. After 365 days: $40 * 365 = 14,600$ MB = 14.6 GB (1 GB = 1,000 MB).
4. Chữ V rõ nhất là tốc độ (velocity): bản ghi đến liên tục suốt ngày. Theo cách tổ chức, đây là chuỗi sự kiện, dạng dữ liệu bán cấu trúc. Dung lượng 14,6 GB mỗi năm vừa một máy tính; "lớn" ở đây nằm ở nhịp đến của dữ liệu hơn là kích thước.
The clearest V is velocity: records arrive continuously all day. By organisation this is an

event stream, semi-structured data. 14.6 GB a year fits on one computer; the "big" here lies in the pace of arrival more than in the size.

1.13. Câu hỏi tự luận và dàn ý trả lời — Essay questions and answer outlines

Câu 1. Phân biệt dữ liệu, thông tin và tri thức; phân biệt KDD với khai phá dữ liệu. Minh họa bằng bảng khách hàng và bảng hồ sơ vay của Phú Sa.

Question 1. *Distinguish data, information and knowledge; distinguish KDD from data mining. Illustrate with Phú Sa's customer table and loan application table.*

- Dữ liệu là giá trị thô (dòng của KH03); thông tin là dữ liệu đã tổng hợp trả lời một câu hỏi (7 trong 12 khách có nhà); tri thức là mẫu khái quát dùng cho dữ liệu mới.
Data is a raw value (the row of KH03); information is summarised data answering a question (7 of 12 customers own a house); knowledge is a general pattern usable on new data.
- Ví dụ tri thức: trong 14 hồ sơ, lịch sử Xấu đi cùng tỷ lệ nợ xấu 80%, lịch sử Tốt đi cùng 0%; quy tắc theo lịch sử đúng 12/14 hồ sơ (85,7%).
A knowledge example: among 14 applications a Bad history goes with an 80% bad-debt rate and a Good history with 0%; the rule by history is right on 12 of 14 (85.7%).
- KDD là cả quá trình năm bước: chọn dữ liệu, làm sạch và tích hợp, biến đổi và thu giảm, khai phá, đánh giá và trình bày.
KDD is the whole five-step process: select data, clean and integrate, transform and reduce, mine, evaluate and present.
- Khai phá dữ liệu theo nghĩa hẹp chỉ là bước thứ tư, chạy thuật toán tìm mẫu; các bước còn lại chuẩn bị và đánh giá.
Data mining in the narrow sense is only the fourth step, running the pattern-finding algorithm; the other steps prepare and evaluate.
- Mẫu có giá trị khi hợp lệ, mới, hữu ích và hiểu được; tri thức phải kiểm trên dữ liệu chưa dùng để lập nó.
A pattern has value when valid, novel, useful and understandable; knowledge must be tested on data not used to form it.

Câu 2. Trình bày sáu pha CRISP-DM áp vào dự án chấm điểm tín dụng của Phú Sa và giải thích vì sao quy trình mang tính lặp.

Question 2. *Describe the six CRISP-DM phases applied to Phú Sa's credit scoring project and explain why the process is iterative.*

- Pha 1 hiểu nghiệp vụ: mục tiêu giảm khoản vay thành nợ xấu mà không từ chối nhiều khách tốt, kèm tiêu chí thành công.
Phase 1 business understanding: the goal of cutting loans that become bad debt without rejecting many good customers, with success criteria.
- Pha 2 hiểu dữ liệu: thống kê thu nhập, dư nợ, phát hiện giá trị thiếu (KH07) và ngoại lai (KH12).
Phase 2 data understanding: summarise income and debt, find the missing value (KH07) and the outlier (KH12).

- Pha 3 chuẩn bị dữ liệu: điền thiếu, tính DTI, chuẩn hóa, chia tập huấn luyện và kiểm tra.

Phase 3 data preparation: fill gaps, compute DTI, normalise, split into training and test sets.

- Pha 4 mô hình hóa, pha 5 đánh giá trên tập kiểm tra theo tiêu chí nghiệp vụ, pha 6 triển khai và giám sát độ ổn định điểm.

Phase 4 modelling, phase 5 evaluation on the test set against the business criteria, phase 6 deployment and monitoring of score stability.

- Quy trình lặp vì kết quả của một pha có thể làm đổi pha trước: dữ liệu không đủ thì đổi mục tiêu, mô hình chưa đạt thì quay lại chuẩn bị dữ liệu.

The process iterates because the result of one phase can change an earlier one: insufficient data changes the goal, a model that falls short sends the team back to data preparation.

Câu 3. So sánh OLAP với khai phá dữ liệu; nêu các thách thức và quy định pháp luật về thông tin khách hàng khi Phú Sa làm dự án chấm điểm tín dụng.

Question 3. Compare OLAP with data mining; state the challenges and the legal rules on customer information when Phú Sa runs a credit scoring project.

- OLAP tổng hợp theo chiều do người dùng chọn, trả lời "bao nhiêu, ở đâu" (dư nợ theo hạng thẻ); khai phá dữ liệu tự tìm mẫu, trả lời "vì sao, ai sẽ".

OLAP aggregates along dimensions the user chooses and answers "how many, where" (debt by card tier); data mining finds patterns itself and answers "why, who will".

- Thách thức: chất lượng dữ liệu, quyền riêng tư, giải thích được, quá khớp, thiên lệch; mỗi thách thức nêu biểu hiện ở Phú Sa và cách xử lý.

Challenges: data quality, privacy, explainability, overfitting, bias; for each give how it shows at Phú Sa and how to handle it.

- Luật Các tổ chức tín dụng (bản hợp nhất 143/2026) Điều 13: nhân viên không tiết lộ thông tin khách hàng; không cung cấp cho tổ chức, cá nhân khác, trừ yêu cầu của cơ quan có thẩm quyền theo luật hoặc khách hàng chấp thuận.

Law on Credit Institutions (consolidated text 143/2026) Article 13: staff may not disclose customer information; no provision to other organisations or individuals except at the request of a competent body under law or with the customer's consent.

- Luật Bảo vệ dữ liệu cá nhân 91/2025 Điều 2 khoản 1, 3, 11 và Nghị định 356/2025 Điều 4 khoản 1 điểm k: thông tin tài chính, tín dụng là dữ liệu nhạy cảm; khử nhận dạng làm dữ liệu không còn xác định được người.

Law on Personal Data Protection 91/2025 Article 2, clauses 1, 3, 11 and Decree 356/2025 Article 4, clause 1(k): financial and credit information is sensitive data; de-identification leaves data unable to identify a person.

- Nghị định 356/2025 Điều 8 khoản 2 điểm a: khi xin đồng ý phải nêu rõ mục đích, gồm chấm điểm, xếp hạng tín dụng nếu có.

Decree 356/2025 Article 8, clause 2(a): when seeking consent the purposes must be stated, including credit scoring and rating where applicable.

1.14. Câu hỏi trắc nghiệm — Multiple-choice questions

1. Trong bảng khách hàng của Phú Sa, phát biểu nào là tri thức theo nghĩa khai phá dữ liệu?

In Phú Sa's customer table, which statement is knowledge in the data mining sense?

- A. Dòng ghi KH03 là 34 tuổi, giáo viên, hạng thẻ Vàng, dư nợ 20 / *The row KH03 records age 34, teacher, Gold tier, debt 20*
- B. Bảng có 12 khách hàng, mỗi nghề có đúng 3 khách hàng / *The table has 12 customers, exactly 3 customers in each occupation*
- C. Tổng dư nợ của 12 khách hàng là 415 triệu đồng / *The total debt of the 12 customers is 415 million dong*
- D. Lịch sử Tốt đi cùng nợ xấu thấp, dùng để sàng hồ sơ mới / *A Good history goes with low bad debt, used to screen new applications*

2. Dựng bảng huấn luyện từ nhiều nguồn, điền giá trị thiếu và chia tập huấn luyện, tập kiểm tra thuộc pha nào của CRISP-DM?

Building the training table from several sources, filling missing values and splitting training and test sets belong to which CRISP-DM phase?

- A. Hiểu nghiệp vụ / *Business understanding*
- B. Hiểu dữ liệu / *Data understanding*
- C. Mô hình hóa / *Modelling*
- D. Chuẩn bị dữ liệu / *Data preparation*

3. Bài toán nào sau đây là phân cụm?

Which of the following problems is clustering?

- A. Dự báo dư nợ cuối tháng 10/2026 từ chuỗi 12 tháng trước / *Forecasting the month-end debt for October 2026 from the previous 12 months*
- B. Chia khách thành nhóm theo số giao dịch và chi tiêu khi chưa có nhóm cho trước / *Dividing customers into groups by transactions and spending when no groups are given*
- C. Đoán hồ sơ vay mới là Tốt hay Xấu từ các hồ sơ đã đáo hạn / *Predicting whether a new application is Good or Bad from matured applications*
- D. Tìm các cặp sản phẩm hay xuất hiện cùng nhau trong giao dịch / *Finding pairs of products that often appear together in transactions*

4. Trong 8 hồ sơ có 3 hồ sơ nợ xấu. Quy tắc luôn đoán "không nợ xấu" có độ chính xác bao nhiêu?

Among 8 applications 3 have bad debt. What is the accuracy of the rule that always predicts "no bad debt"?

- A. 37,5% / *37.5%*
- B. 50%
- C. 62,5% / *62.5%*
- D. 75%

5. Tập dữ liệu có cột nhãn nợ xấu đã biết; mục tiêu là đoán nhãn cho hồ sơ mới. Đây là kiểu học nào?
A data set has a known bad-debt label column; the goal is to predict the label of new applications. Which type of learning is this?
- A. Học không giám sát / *Unsupervised learning*
 - B. Học bán giám sát / *Semi-supervised learning*
 - C. Phân cụm / *Clustering*
 - D. Học có giám sát / *Supervised learning*
6. Mỗi ngày ứng dụng ghi 10.000 sự kiện, mỗi sự kiện 300 byte (1 MB = 1.000.000 byte). Dung lượng sau 30 ngày là
Each day the app records 10,000 events of 300 bytes each (1 MB = 1,000,000 bytes). The storage after 30 days is
- A. 9 MB
 - B. 30 MB
 - C. 90 MB
 - D. 900 MB
7. Mã Python "print(sum(x > 40 for x in [52, 46, 40, 30, 44]))" in ra giá trị nào?
Which value does the Python code "print(sum(x > 40 for x in [52, 46, 40, 30, 44]))" print?
- A. 2
 - B. 4
 - C. 3
 - D. 5
8. Theo Luật Bảo vệ dữ liệu cá nhân 91/2025 Điều 2 khoản 1, dữ liệu cá nhân sau khi khử nhận dạng thì
Under the Law on Personal Data Protection 91/2025, Article 2, clause 1, personal data after de-identification
- A. vẫn là dữ liệu cá nhân nhạy cảm / *remains sensitive personal data*
 - B. không còn là dữ liệu cá nhân / *is no longer personal data*
 - C. phải được xóa ngay lập tức / *must be deleted immediately*
 - D. trở thành bí mật nhà nước / *becomes a state secret*

Chương 2. Dữ liệu, thống kê mô tả và độ đo tương tự — Chapter 2. Data, descriptive statistics and similarity measures

Mục tiêu chương — Chapter objectives

- Phân loại thuộc tính theo kiểu (định danh, nhị phân, thứ bậc, số khoảng, số tỷ lệ) và theo tập giá trị (rời rạc, liên tục), rồi chọn phép toán hợp lệ cho từng kiểu.
Classify attributes by type (nominal, binary, ordinal, interval, ratio) and by value set (discrete, continuous), then choose the valid operations for each type.
- Tính tay trung bình, trung vị, mode, phương sai, độ lệch chuẩn, tứ phân vị, IQR và năm số tóm tắt của thu nhập khách hàng Phú Sa, phát hiện ngoại lai theo quy tắc $1.5 * IQR$.
Compute by hand the mean, median, mode, variance, standard deviation, quartiles, IQR and five-number summary of Phú Sa customer income, and detect outliers with the $1.5 * IQR$ rule.
- Đọc hình hộp, histogram và độ lệch; tính hệ số tương quan Pearson giữa thu nhập và dư nợ và nêu giới hạn của nó.
Read boxplots, histograms and skewness; compute the Pearson correlation between income and balance and state its limits.
- Lập ma trận khoảng cách sau chuẩn hóa min–max và tính khoảng cách Minkowski (Manhattan, Euclid, Chebyshev); kiểm tính chất metric.
Build a dissimilarity matrix after min–max normalisation and compute Minkowski distances (Manhattan, Euclid, Chebyshev); test the metric properties.
- Tính SMC, Jaccard, cosine và khoảng cách Gower cho dữ liệu nhị phân, định danh, thứ bậc và hỗn hợp.
Compute SMC, Jaccard, cosine and the Gower distance for binary, nominal, ordinal and mixed data.

2.1. Đối tượng, thuộc tính và kiểu thuộc tính — 2.1. Objects, attributes and attribute types

Một tập dữ liệu gồm các **đối tượng** (object; còn gọi là bản ghi, mẫu, điểm dữ liệu). Mỗi đối tượng được mô tả bằng một tập **thuộc tính** (attribute; còn gọi là biến, trường, cột, đặc trưng). Ở Công ty Tài chính TNHH Phú Sa, mỗi khách hàng là một đối tượng, gọi bằng mã KH01 đến KH12; mỗi khách được mô tả bằng bảy thuộc tính: tuổi, thu nhập, nghề, hạng thẻ, dư nợ, có nhà và số lần trễ hạn. Trong cơ sở dữ liệu quan hệ của học phần Cơ sở dữ liệu, đối tượng là một dòng (bộ), thuộc tính là một cột, và các hàm AVG, COUNT, MIN, MAX của SQL cho cùng kết quả với các độ đo ở mục 2.2 và 2.3.

*A data set consists of **objects** (also called records, samples or data points). Each object is described by a set of **attributes** (also called variables, fields, columns or features). At Phú Sa Finance Company Ltd, each customer is an object, named KH01 to KH12, and each customer is described by seven attributes: age, income, occupation, card tier, outstanding balance, home ownership and number of late payments. In the relational database of the Databases course an*

object is a row (tuple) and an attribute is a column, and the SQL functions AVG, COUNT, MIN and MAX give the same results as the measures in sections 2.2 and 2.3.

Mã ID	Tuổi Age	Thu nhập Income	Nghề Occupation	Hạng thẻ Card tier	Dư nợ Balance	Có nhà Home	Trễ hạn Late
KH01	25	9	Công nhân Factory worker	Chuẩn Standard	12	0	2
KH02	29	12	Nhân viên văn phòng Office clerk	Chuẩn Standard	18	0	1
KH03	34	15	Giáo viên Teacher	Vàng Gold	20	1	0
KH04	41	22	Kinh doanh tự do Self-employed	Vàng Gold	45	1	1
KH05	23	8	Công nhân Factory worker	Chuẩn Standard	10	0	4
KH06	38	18	Nhân viên văn phòng Office clerk	Vàng Gold	30	1	0
KH07	45	—	Kinh doanh tự do Self-employed	Vàng Gold	40	1	2
KH08	31	14	Giáo viên Teacher	Chuẩn Standard	16	0	0
KH09	27	11	Nhân viên văn phòng Office clerk	Chuẩn Standard	14	0	1
KH10	36	16	Công nhân Factory worker	Vàng Gold	25	1	3
KH11	50	20	Giáo viên Teacher	Bạch kim Platinum	35	1	0
KH12	47	120	Kinh doanh tự do Self-employed	Bạch kim Platinum	150	1	0

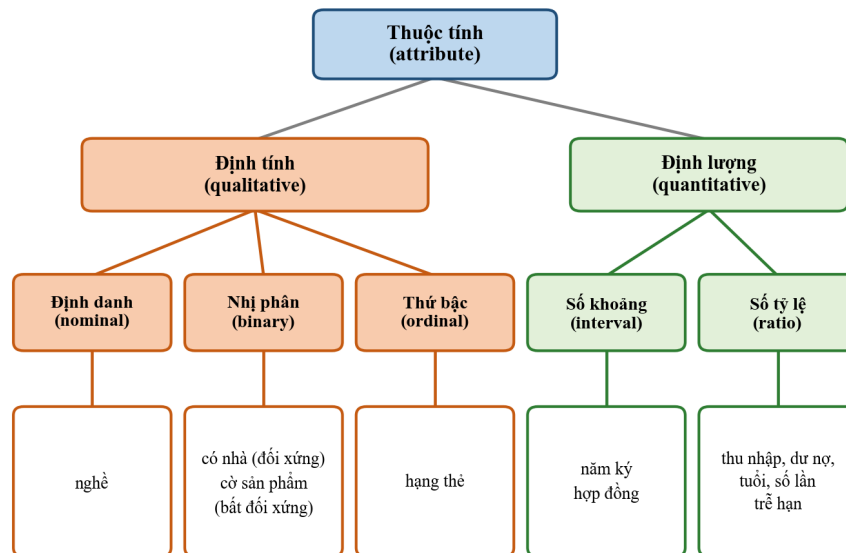
Thu nhập tính bằng triệu đồng mỗi tháng, dư nợ bằng triệu đồng; cột có nhà ghi 1 nếu khách sở hữu nhà, 0 nếu không. Ô thu nhập của KH07 bị thiếu (dấu gạch), nên các thống kê về thu nhập tính trên 11 khách; cách xử lý giá trị thiếu thuộc Chương 3. Khách KH12 có thu nhập 120 triệu đồng, gấp 8 lần trung vị, và sẽ là ngoại lai điển hình ở mục 2.3.

Income is in VND million per month and balance in VND million; the home column is 1 if the customer owns a home and 0 otherwise. The income cell of KH07 is missing (dash), so income statistics are computed over 11 customers; handling missing values belongs to Chapter 3. Customer KH12 earns VND 120 million, 8 times the median, and is the typical outlier of section 2.3.

Kiểu thuộc tính — *Attribute types*

Kiểu của thuộc tính quyết định phép toán nào có nghĩa: tính trung bình mã nghề là vô nghĩa, còn tính trung bình dư nợ thì có nghĩa. Có hai nhóm lớn, mỗi nhóm chia nhỏ như Hình 2.1: thuộc tính định tính (giá trị là nhãn) và thuộc tính định lượng (giá trị là số đo).

The type of an attribute decides which operations are meaningful: the average of an occupation code is meaningless, while the average balance is meaningful. There are two broad groups, subdivided as in Figure 2.1: qualitative attributes (values are labels) and quantitative attributes (values are measurements).



Ồ dưới mỗi kiểu: ví dụ lấy từ bảng khách hàng của Phú Sa

Hình 2.1. Phân loại thuộc tính, ví dụ lấy từ bảng khách hàng của Phú Sa

Figure 2.1. Classification of attributes, with examples from the Phú Sa customer table

Kiểu <i>Type</i>	Đặc điểm <i>Characteristics</i>	Phép có nghĩa <i>Meaningful operations</i>	Ví dụ ở Phú Sa <i>Phú Sa example</i>
Định danh <i>Nominal</i>	Giá trị là tên hoặc nhãn, không có thứ tự <i>Values are names or labels with no order</i>	So sánh bằng hoặc khác; đếm; mode <i>Equal or not equal; counting; mode</i>	Nghề: Giáo viên, Công nhân <i>Occupation: Teacher, Factory worker</i>
Nhị phân đối xứng <i>Symmetric binary</i>	Hai giá trị, hai giá trị có tầm quan trọng như nhau <i>Two values of equal importance</i>	Như định danh <i>As for nominal</i>	Có nhà: 1 hoặc 0 <i>Home ownership: 1 or 0</i>
Nhị phân bất đối xứng <i>Asymmetric binary</i>	Hai giá trị, giá trị 1 (có xuất hiện) mới mang thông tin <i>Two values, only 1 (present) carries information</i>	Như định danh; so sánh tập hợp (Jaccard) <i>As for nominal; set comparison (Jaccard)</i>	Cờ dùng bảo hiểm khoản vay <i>Flag for using loan insurance</i>

Kiểu Type	Đặc điểm Characteristics	Phép có nghĩa Meaningful operations	Ví dụ ở Phú Sa Phú Sa example
Thứ bậc Ordinal	Có thứ tự, khoảng cách giữa hai bậc không xác định <i>Ordered, gap between ranks undefined</i>	Thêm so sánh lớn hơn, nhỏ hơn; trung vị; tứ phân vị <i>Adds greater or less than; median; quartiles</i>	Hạng thẻ: Chuẩn < Vàng < Bạch kim <i>Card tier: Standard < Gold < Platinum</i>
Số khoảng Interval	Khoảng cách có nghĩa, không có điểm không tuyệt đối <i>Differences are meaningful, no true zero</i>	Thêm cộng, trừ; trung bình; độ lệch chuẩn; tỷ số vô nghĩa <i>Adds addition, subtraction, mean, standard deviation; ratios are meaningless</i>	Năm ký hợp đồng: 2019, 2024 <i>Contract year: 2019, 2024</i>
Số tỷ lệ Ratio	Có điểm không tuyệt đối, tỷ số có nghĩa <i>True zero, ratios are meaningful</i>	Mọi phép toán, kể cả tỷ số <i>All operations, including ratios</i>	Thu nhập 18 triệu gấp đôi 9 triệu <i>Income of 18 million is twice 9 million</i>

Cách phân chia khác theo tập giá trị: thuộc tính **rời rạc** (discrete) nhận hữu hạn hoặc đếm được giá trị, như số lần trễ hạn (0, 1, 2, 3, 4), hạng thẻ, nghề; thuộc tính **liên tục** (continuous) nhận giá trị thực trong một khoảng, như thu nhập, dư nợ. Thuộc tính nhị phân là trường hợp riêng của rời rạc. Dữ liệu máy tính chỉ lưu số có hữu hạn chữ số thập phân, nên liên tục là một mô hình của thuộc tính. Chương trình sau khai báo lại bảng khách hàng và đếm số giá trị khác nhau, số ô thiếu của từng thuộc tính.

*A second split is by the set of values: a **discrete** attribute takes finitely or countably many values, such as the number of late payments (0, 1, 2, 3, 4), card tier or occupation; a **continuous** attribute takes real values in an interval, such as income or balance. A binary attribute is a special case of a discrete one. A computer stores only numbers with finitely many decimals, so continuous is a model of the attribute. The following program declares the customer table again and counts the distinct values and the missing cells of each attribute.*

```
khach = [
    ("KH01", 25, 9, "Công nhân", "Chuẩn", 12, 0, 2),
    ("KH02", 29, 12, "Nhân viên văn phòng", "Chuẩn", 18, 0, 1),
    ("KH03", 34, 15, "Giáo viên", "Vàng", 20, 1, 0),
    ("KH04", 41, 22, "Kinh doanh tự do", "Vàng", 45, 1, 1),
    ("KH05", 23, 8, "Công nhân", "Chuẩn", 10, 0, 4),
    ("KH06", 38, 18, "Nhân viên văn phòng", "Vàng", 30, 1, 0),
    ("KH07", 45, None, "Kinh doanh tự do", "Vàng", 40, 1, 2),
    ("KH08", 31, 14, "Giáo viên", "Chuẩn", 16, 0, 0),
    ("KH09", 27, 11, "Nhân viên văn phòng", "Chuẩn", 14, 0, 1),
    ("KH10", 36, 16, "Công nhân", "Vàng", 25, 1, 3),
    ("KH11", 50, 20, "Giáo viên", "Bạch kim", 35, 1, 0),
    ("KH12", 47, 120, "Kinh doanh tự do", "Bạch kim", 150, 1, 0),
]
ten = ["ma", "tuoi", "thu_nhap", "nghe", "hang_the", "du_no",
       "co_nha", "so_lan_tru_han"]
for j, t in enumerate(ten):
    cot = [dong[j] for dong in khach]
    co_gia_tri = [x for x in cot if x is not None]
    print(f"{t:<15}{len(set(co_gia_tri)):>3} giá trị khác nhau, "
          f"thiếu {len(cot) - len(co_gia_tri)}")
```

Kết quả chạy / Output

ma	12 giá trị khác nhau, thiếu 0
tuoi	12 giá trị khác nhau, thiếu 0
thu_nhap	11 giá trị khác nhau, thiếu 1
nghe	4 giá trị khác nhau, thiếu 0
hang_the	3 giá trị khác nhau, thiếu 0
du_no	12 giá trị khác nhau, thiếu 0
co_nha	2 giá trị khác nhau, thiếu 0
so_lan_tre_han	5 giá trị khác nhau, thiếu 0

Thu nhập có 11 giá trị khác nhau và thiếu 1 ô; nghề có 4 giá trị, hạng thẻ có 3, số lần trễ hạn có 5. Mã khách là định danh với 12 giá trị khác nhau, nên không đưa vào tính toán.

Income has 11 distinct values and 1 missing cell; occupation has 4 values, card tier 3 and late payments 5. The customer code is a nominal identifier with 12 distinct values and is left out of calculations.

PHẢI THUỘC / MEMORISE — Định danh chỉ đếm; thứ bậc thêm trung vị; số khoảng thêm trung bình; số tỷ lệ thêm tỷ số. Nhị phân bất đối xứng dùng Jaccard, nhị phân đối xứng dùng khớp đơn giản (mục 2.7).

Nominal only counts; ordinal adds the median; interval adds the mean; ratio adds ratios. Asymmetric binary uses Jaccard, symmetric binary uses simple matching (section 2.7).

LỖI HAY GẶP / COMMON MISTAKE — Mã hóa hạng thẻ thành 1, 2, 3 rồi coi như số tỷ lệ: trung bình hạng 2,0 không có nghĩa là hạng Vàng, và hạng Bạch kim không phải gấp ba hạng Chuẩn. Cũng không tính trung bình mã khách hàng hay số điện thoại.

Coding card tier as 1, 2, 3 and treating it as a ratio attribute: an average rank of 2.0 does not mean Gold, and Platinum is not three times Standard. Do not average customer codes or phone numbers either.

2.2. Đo xu hướng trung tâm — 2.2. Measures of central tendency

Độ đo xu hướng trung tâm trả lời câu hỏi giá trị điển hình của thuộc tính là bao nhiêu. Với n giá trị $x_1, \dots, x_n$, sắp xếp tăng dần thành $x(1) \leq \dots \leq x(n)$, các độ đo gồm:

A measure of central tendency answers what the typical value of an attribute is. For n values $x_1, \dots, x_n$ sorted into $x(1) \leq \dots \leq x(n)$, the measures are:

$$\bar{x} = (\sum x_i) / n \quad (2.1)$$

- **Trung bình (mean)** cộng hết rồi chia n , công thức (2.1); dùng được cho số khoảng và số tỷ lệ.

The mean adds everything and divides by n , formula (2.1); usable for interval and ratio attributes.

$$\bar{x}_\alpha = (\sum x_{(i)}) / (n - 2k), i = k + 1, \dots, n - k; k = \lfloor \alpha * n \rfloor \quad (2.2)$$

- **Trung bình cắt tỉa (trimmed mean)** bỏ k giá trị nhỏ nhất và k giá trị lớn nhất rồi lấy trung bình, công thức (2.2). Sách này cắt tỉa 10% mỗi đầu ($\alpha = 0,1$).

The trimmed mean drops the k smallest and k largest values and averages the rest, formula (2.2). This book trims 10% at each end ($\alpha = 0.1$).

$$\text{trung vị} = x_{((n+1)/2)} \text{ khi } n \text{ lẻ; trung vị} = (x_{(n/2)} + x_{(n/2+1)}) / 2 \text{ khi } n \text{ chẵn} \quad (2.3)$$

- **Trung vị** (median) là giá trị ở giữa sau khi sắp xếp, công thức (2.3); dùng được từ thuộc tính thứ bậc trở lên.
*The **median** is the middle value after sorting, formula (2.3); usable from ordinal attributes upward.*
- **Mode** là giá trị xuất hiện nhiều nhất; có thể có nhiều mode hoặc không có mode nào; dùng được cho mọi kiểu thuộc tính.
*The **mode** is the most frequent value; there may be several modes or none; usable for every attribute type.*

$$\text{midrange} = (x_{\min} + x_{\max}) / 2 \quad (2.4)$$

- **Midrange** là trung điểm của giá trị nhỏ nhất và lớn nhất, công thức (2.4); tính nhanh nhưng bị ngoại lai chi phối.
*The **midrange** is the midpoint of the smallest and largest values, formula (2.4); quick to compute but dominated by outliers.*

```
import statistics as st
from collections import Counter

thu_nhap = [9, 12, 15, 22, 8, 18, 14, 11, 16, 20, 120] # bỏ KH07
tre_han = [2, 1, 0, 1, 4, 0, 2, 0, 1, 3, 0, 0] # cả 12 khách
n = len(thu_nhap)
x = sorted(thu_nhap)
k = int(0.1 * n) # cắt tía 10% mỗi đầu
print("n =", n, "| tổng =", sum(x))
print("trung bình =", round(st.mean(x), 2))
print("trung bình cắt tía 10% =", round(st.mean(x[k:n - k]), 2))
print("trung vị =", st.median(x))
print("midrange =", (x[0] + x[-1]) / 2)
print("thu nhập, giá trị gặp nhiều nhất:", Counter(x).most_common(1)[0])
mode, so_lan = Counter(tre_han).most_common(1)[0]
print("số lần trễ hạn: mode =", mode, "xuất hiện", so_lan, "lần")
```

Kết quả chạy / Output

```
n = 11 | tổng = 265
trung bình = 24.09
trung bình cắt tía 10% = 15.22
trung vị = 15
midrange = 64.0
thu nhập, giá trị gặp nhiều nhất: (8, 1)
số lần trễ hạn: mode = 0 xuất hiện 5 lần
```

Trung bình thu nhập là 24,09 triệu đồng, còn trung vị chỉ 15 triệu đồng: KH12 kéo trung bình lên 1,6 lần trung vị. Trung bình cắt tía 10% là 15,22, gần trung vị vì đã bỏ KH05 và KH12 (hai đầu). Midrange bằng 64 triệu đồng, không đại diện cho khách nào. Thu nhập không có mode vì mỗi giá trị chỉ xuất hiện một lần; số lần trễ hạn có mode 0 (5 khách). Với nghề, bốn nghề đều xuất hiện đúng 3 lần, nên tập dữ liệu có bốn mode; bộ đếm Counter của Python trả về Công nhân chỉ vì đó là nghề gặp đầu tiên trong danh sách.

Mean income is 24.09 million but the median is only 15 million: KH12 pulls the mean up to 1.6 times the median. The 10% trimmed mean is 15.22, close to the median because both ends (KH05 and KH12) are dropped. The midrange is 64 million and represents no customer. Income has no mode because every value occurs once; late payments have mode 0 (5 customers). For occupation all four occupations occur exactly 3 times, so the data set has four modes; the Python Counter returns Factory worker only because it is the first occupation met in the list.

PHẢI THUỘC / MEMORISE — Ngoại lai kéo trung bình và midrange lệch nhiều, còn trung vị gần như đứng yên. Phân phối đối xứng thì trung bình, trung vị trùng nhau; lệch phải thì trung bình lớn hơn trung vị.

Outliers drag the mean and the midrange far, and barely move the median. A symmetric distribution has mean equal to median; a right-skewed one has mean above median.

Bài mẫu 2.1. Trung bình, trung vị, cắt tỉa của dư nợ — *Worked example 2.1. Mean, median and trimmed mean of balance*

Đề bài: Bảng khách hàng có 12 giá trị dư nợ (triệu đồng): 12, 18, 20, 45, 10, 30, 40, 16, 14, 25, 35, 150. Tính trung bình, trung vị, trung bình cắt tỉa 10% và midrange.

Problem: the customer table has 12 balances (VND million): 12, 18, 20, 45, 10, 30, 40, 16, 14, 25, 35, 150. Compute the mean, the median, the 10% trimmed mean and the midrange.

Thứ tự Rank	Khách Customer	Dư nợ Balance	Thứ tự Rank	Khách Customer	Dư nợ Balance
1	KH05	10	7	KH10	25
2	KH01	12	8	KH06	30
3	KH09	14	9	KH11	35
4	KH08	16	10	KH07	40
5	KH02	18	11	KH04	45
6	KH03	20	12	KH12	150

1. Trung bình: tổng = 415, $n = 12$, nên trung bình = $415 / 12 = 34,58$ triệu đồng.
Mean: sum = 415, $n = 12$, so the mean = $415 / 12 = 34.58$ million.
2. Trung vị: $n = 12$ chẵn, lấy trung bình hai giá trị ở vị trí 6 và 7 sau khi sắp xếp: $(20 + 25) / 2 = 22,5$.
Median: $n = 12$ is even, so average the values at positions 6 and 7 after sorting: $(20 + 25) / 2 = 22.5$.
3. Cắt tỉa 10%: $k = \lfloor 0,1 * 12 \rfloor = 1$, bỏ giá trị nhỏ nhất (10) và lớn nhất (150); trung bình 10 giá trị còn lại = $(415 - 10 - 150) / 10 = 25,5$.
10% trimmed mean: $k = \lfloor 0.1 * 12 \rfloor = 1$, drop the smallest (10) and the largest (150); mean of the remaining 10 values = $(415 - 10 - 150) / 10 = 25.5$.
4. Midrange = $(10 + 150) / 2 = 80$.
Midrange = $(10 + 150) / 2 = 80$.
5. Nhận xét: trung bình 34,58 lớn hơn trung vị 22,5 vì KH12 có dư nợ 150; bỏ hai đầu thì trung bình về 25,5. Dư nợ không có mode.
Comment: the mean 34.58 exceeds the median 22.5 because KH12 owes 150; dropping both ends brings the mean to 25.5. Balance has no mode.

2.3. Đo độ phân tán và năm số tóm tắt — 2.3. Measures of dispersion and the five-number summary

Hai tập thu nhập có cùng trung bình vẫn có thể rải rộng khác nhau, nên cần đo độ phân tán. Khoảng biến thiên (range) là hiệu giữa lớn nhất và nhỏ nhất. Phương sai đo trung bình bình phương độ lệch so với trung bình; độ lệch chuẩn là căn bậc hai của phương sai và có cùng đơn vị với dữ liệu.

Two income sets with the same mean can still be spread differently, so a measure of dispersion is needed. The range is the difference between the largest and smallest values. The variance is the mean squared deviation from the mean; the standard deviation is its square root and has the same unit as the data.

$$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1); \sigma^2 = \Sigma (x_i - \mu)^2 / n; s = \sqrt{s^2} \quad (2.5)$$

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n \quad (2.6)$$

Đặc điểm <i>Feature</i>	Phương sai mẫu s^2 <i>Sample variance s^2</i>	Phương sai tổng thể σ^2 <i>Population variance σ^2</i>
Mẫu số <i>Denominator</i>	$n - 1$	n
Dùng khi <i>Use when</i>	Dữ liệu là mẫu lấy từ tổng thể lớn hơn <i>The data are a sample from a larger population</i>	Dữ liệu là toàn bộ tổng thể <i>The data are the whole population</i>
Hàm Python <i>Python function</i>	statistics.variance, statistics.stdev	statistics.pvariance, statistics pstdev
Mặc định của thư viện <i>Library defaults</i>	pandas (ddof = 1), Excel VAR.S	numpy.var (ddof = 0), Excel VAR.P

Sách này dùng phương sai và độ lệch chuẩn **mẫu** (mẫu số $n - 1$) cho mọi chương, trừ khi đề bài ghi rõ là tổng thể. Công thức (2.6) cho phép tính tay mà không cần từng độ lệch: chỉ cần tổng các giá trị và tổng bình phương.

*This book uses the **sample** variance and standard deviation (denominator $n - 1$) in every chapter, unless the problem says population. Formula (2.6) allows hand computation without each deviation: only the sum of values and the sum of squares are needed.*

```
import statistics as st

thu_nhap = [9, 12, 15, 22, 8, 18, 14, 11, 16, 20, 120]
n = len(thu_nhap)
tb = sum(thu_nhap) / n
tong_bp = sum((x - tb) ** 2 for x in thu_nhap)
print("khoảng biến thiên =", max(thu_nhap) - min(thu_nhap))
print("tổng bình phương lệch =", round(tong_bp, 2))
print("phương sai mẫu =", round(tong_bp / (n - 1), 2))
print("phương sai tổng thể =", round(tong_bp / n, 2))
print("độ lệch chuẩn mẫu =", round(st.stdev(thu_nhap), 2))
print("độ lệch chuẩn tổng thể =", round(st.pstdev(thu_nhap), 2))
print("hệ số biến thiên =", round(st.stdev(thu_nhap) / tb * 100, 1), "%")
```

Kết quả chạy / Output

```

khoảng biến thiên = 112
tổng bình phương lệch = 10310.91
phương sai mẫu = 1031.09
phương sai tổng thể = 937.36
độ lệch chuẩn mẫu = 32.11
độ lệch chuẩn tổng thể = 30.62
hệ số biến thiên = 133.3 %

```

Thu nhập có khoảng biến thiên 112, phương sai mẫu 1.031,09, độ lệch chuẩn mẫu 32,11 triệu đồng. Hệ số biến thiên (độ lệch chuẩn chia trung bình) là 133,3%, nghĩa là độ rải lớn hơn cả mức trung bình, một dấu hiệu của ngoại lai. Phương sai tổng thể nhỏ hơn (937,36) vì chia cho n .

Income has a range of 112, a sample variance of 1,031.09 and a sample standard deviation of 32.11 million. The coefficient of variation (standard deviation divided by the mean) is 133.3%, meaning the spread exceeds the mean itself, a sign of outliers. The population variance is smaller (937.36) because it divides by n .

Tứ phân vị, IQR và năm số tóm tắt — Quartiles, IQR and the five-number summary

Tứ phân vị chia dữ liệu đã sắp xếp thành bốn phần có số phần tử xấp xỉ bằng nhau: Q1 (25%), Q2 (trung vị), Q3 (75%). Hiệu IQR = Q3 – Q1 là độ rộng của nửa giữa dữ liệu. Năm số tóm tắt gồm giá trị nhỏ nhất, Q1, trung vị, Q3 và giá trị lớn nhất.

Quartiles split the sorted data into four parts with roughly equal counts: Q1 (25%), Q2 (median) and Q3 (75%). The difference IQR = Q3 – Q1 is the width of the middle half of the data. The five-number summary lists the minimum, Q1, the median, Q3 and the maximum.

CẢNH BÁO / WARNING — Quy ước tứ phân vị của sách (cố định cho mọi chương): vị trí của phân vị p là $1 + (n - 1) * p$ trong dãy đã sắp xếp, nội suy tuyến tính giữa hai giá trị kề; trùng với `statistics.quantiles(method="inclusive")` của Python và hàm `QUARTILE.INC` của Excel. Giáo trình khác có thể dùng quy ước khác và ra Q1, Q3 khác; đề thi cần ghi rõ quy ước.

Quartile convention of this book (fixed for all chapters): the position of percentile p is $1 + (n - 1) * p$ in the sorted list, with linear interpolation between neighbours; it matches Python `statistics.quantiles(method="inclusive")` and Excel `QUARTILE.INC`. Other textbooks may use another convention and obtain different Q1 and Q3; an exam should state the convention.

vị trí = $1 + (n - 1) * p$; $Q = x_{(j)} + f * (x_{(j+1)} - x_{(j)})$, với j là phần nguyên, f là phần thập phân của vị trí (2.7)

IQR = Q3 – Q1; hàng rào dưới = Q1 – 1,5 * IQR; hàng rào trên = Q3 + 1,5 * IQR (2.8)

Giá trị nằm ngoài hai hàng rào ở công thức (2.8) bị coi là ngoại lai theo quy tắc 1,5 * IQR của Tukey. Cũng có thể kiểm tra bằng z-score (bài toán nâng cao ở Chương 9).

*A value outside the two fences of formula (2.8) is treated as an outlier under Tukey's 1.5 * IQR rule. A z-score check is also possible (an advanced topic in Chapter 9).*

```
import statistics as st

thu_nhap = [9, 12, 15, 22, 8, 18, 14, 11, 16, 20, 120]
x = sorted(thu_nhap)
q1, q2, q3 = st.quantiles(x, n=4, method="inclusive")
iqr = q3 - q1
thap, cao = q1 - 1.5 * iqr, q3 + 1.5 * iqr
print("năm số:", x[0], q1, q2, q3, x[-1])
print("IQR =", iqr, "| hàng rào:", thap, cao)
ngoai_lai = [v for v in x if v < thap or v > cao]
print("ngoại lai:", ngoai_lai)
con_lai = [v for v in x if v not in ngoai_lai]
print("trung bình bỏ ngoại lai =", round(st.mean(con_lai), 2))
print("quy ước exclusive:", st.quantiles(x, n=4, method="exclusive"))
```

Kết quả chạy / Output

```
năm số: 8 11.5 15.0 19.0 120
IQR = 7.5 | hàng rào: 0.25 30.25
ngoại lai: [120]
trung bình bỏ ngoại lai = 14.5
quy ước exclusive: [11.0, 15.0, 20.0]
```

Năm số tóm tắt của thu nhập là 8; 11,5; 15; 19; 120. IQR = 7,5; hàng rào là 0,25 và 30,25, nên chỉ KH12 (120) là ngoại lai. Bỏ KH12, thu nhập trung bình còn 14,5 triệu đồng, sát trung vị 15. Với quy ước exclusive (mẫu số $n + 1$), Q1 = 11 và Q3 = 20, IQR = 9: kết luận ngoại lai vẫn như cũ nhưng hàng rào đổi, vì vậy phải ghi rõ quy ước.

The five-number summary of income is 8, 11.5, 15, 19, 120. IQR = 7.5; the fences are 0.25 and 30.25, so only KH12 (120) is an outlier. Without KH12 the mean income falls to 14.5 million, close to the median 15. Under the exclusive convention (denominator $n + 1$), Q1 = 11 and Q3 = 20, IQR = 9: the outlier verdict is unchanged but the fences move, which is why the convention must be stated.

Bài mẫu 2.2. Phương sai và độ lệch chuẩn của thu nhập — *Worked example 2.2.*

Variance and standard deviation of income

Đề bài: Tính phương sai mẫu và độ lệch chuẩn mẫu của 11 giá trị thu nhập (bỏ KH07), dùng công thức (2.6).

Problem: compute the sample variance and sample standard deviation of the 11 income values (KH07 excluded), using formula (2.6).

Khách <i>Customer</i>	x_i	x_i^2	Khách <i>Customer</i>	x_i	x_i^2
KH01	9	81	KH08	14	196
KH02	12	144	KH09	11	121
KH03	15	225	KH10	16	256
KH04	22	484	KH11	20	400
KH05	8	64	KH12	120	14400
KH06	18	324	Cộng <i>Sum</i>	265	16.695

1. Từ bảng: $n = 11$, $\sum x_i = 265$, $\sum x_i^2 = 16.695$.
From the table: $n = 11$, $\sum x_i = 265$, $\sum x_i^2 = 16,695$.
2. Tổng bình phương lệch: $\sum x_i^2 - (\sum x_i)^2 / n = 16.695 - 265^2 / 11 = 16.695 - 6.384,09 = 10.310,91$.
Sum of squared deviations: $\sum x_i^2 - (\sum x_i)^2 / n = 16,695 - 265^2 / 11 = 16,695 - 6,384.09 = 10,310.91$.
3. Phương sai mẫu: $s^2 = 10.310,91 / (11 - 1) = 1.031,09$.
Sample variance: $s^2 = 10,310.91 / (11 - 1) = 1,031.09$.
4. Độ lệch chuẩn mẫu: $s = \sqrt{1.031,09} = 32,11$ triệu đồng.
Sample standard deviation: $s = \sqrt{1,031.09} = 32.11$ million.

Bài mẫu 2.3. Năm số tóm tắt và ngoại lai của thu nhập — *Worked example 2.3. Five-number summary and outliers of income*

Đề bài: Tìm năm số tóm tắt, IQR và các ngoại lai của 11 giá trị thu nhập theo quy ước của sách.

Problem: find the five-number summary, the IQR and the outliers of the 11 income values under the book's convention.

Vị trí Position	1	2	3	4	5	6	7	8	9	10	11
Giá trị Value	8	9	11	12	14	15	16	18	20	22	120

1. Sắp xếp tăng dần (bảng trên). Giá trị nhỏ nhất 8, lớn nhất 120.
Sort ascending (table above). The minimum is 8 and the maximum is 120.
2. Q1: vị trí = $1 + 10 * 0,25 = 3,5$, giữa giá trị thứ 3 (11) và thứ 4 (12): $Q1 = 11 + 0,5 * (12 - 11) = 11,5$.
Q1: position = $1 + 10 * 0.25 = 3.5$, between the 3rd (11) and 4th (12) values: $Q1 = 11 + 0.5 * (12 - 11) = 11.5$.
3. Trung vị: vị trí = $1 + 10 * 0,5 = 6$, giá trị thứ 6 là 15.
Median: position = $1 + 10 * 0.5 = 6$, and the 6th value is 15.
4. Q3: vị trí = $1 + 10 * 0,75 = 8,5$, giữa giá trị thứ 8 (18) và thứ 9 (20): $Q3 = 18 + 0,5 * (20 - 18) = 19$.
Q3: position = $1 + 10 * 0.75 = 8.5$, between the 8th (18) and 9th (20) values: $Q3 = 18 + 0.5 * (20 - 18) = 19$.
5. $IQR = 19 - 11,5 = 7,5$; $1,5 * IQR = 11,25$. Hàng rào dưới = $11,5 - 11,25 = 0,25$; hàng rào trên = $19 + 11,25 = 30,25$.
 $IQR = 19 - 11.5 = 7.5$; $1.5 * IQR = 11.25$. Lower fence = $11.5 - 11.25 = 0.25$; upper fence = $19 + 11.25 = 30.25$.
6. Giá trị lớn nhất nằm trong hàng rào là 22; giá trị 120 vượt 30,25 nên là ngoại lai duy nhất.
The largest value inside the fences is 22; the value 120 exceeds 30.25 and is the only outlier.