

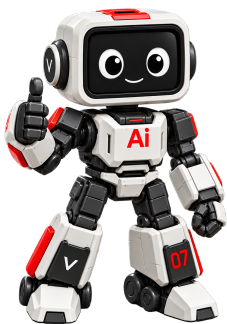


AI IN 7 DAYS

DAY 4 HANDBOOK - SELF-STUDY GUIDE

Major Machine Learning Methods, Forecasting and Evaluation

6 VIDEOS	60 MINUTES	6 REAL-WORLD CASES	1 SELECTION BOARD
-------------	---------------	-----------------------	----------------------



© 2026 AI PANTHIYA

අද ඔබ ඉගෙනගන්නේ

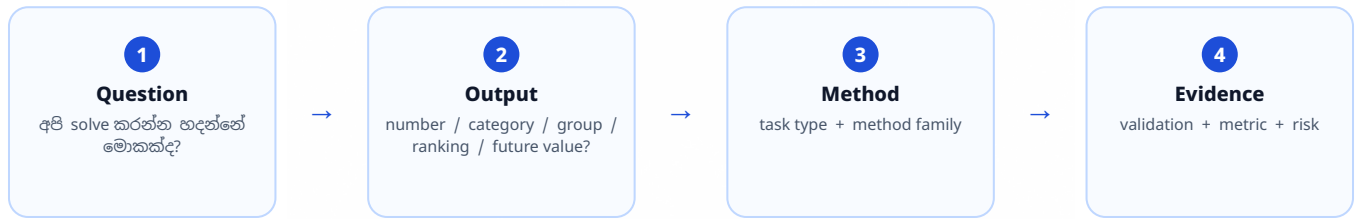
- Problem එකට ගැළපෙන method family එක තෝරන විදිහ
- Clustering සහ hidden groups
- Dimensionality Reduction සහ Anomaly Detection
- Recommendation, Association Rules, Knowledge Graphs
- Forecasting සහ Causal Questions අතර වෙනස
- Task-specific evaluation, robustness සහ explainability



Day 4 Main Task: real-world cases හයක් සඳහා **AI Problem Selection Board** එකක් සකස් කිරීම - Question, Output, Task Type, Method Family, Validation, Metric, Risk සහ Human Review එකට map කිරීම.

HOW TO USE THIS HANDBOOK

Day 4 එකේ main goal එක “popular algorithm එකක් තෝරන එක” තෙවෙයි. Problem එකේ structure, output, available data, chronology, risk සහ evaluation need එක අනුව නිවැරදි method family එක තෝරන reasoning map එකක් ගොඩනගා ගැනීමයි.



Day 4 Learning Outcomes

- Classification, Regression, Ranking, Clustering, Anomaly Detection, Recommendation, Association Rules, Forecasting, Optimisation සහ Causal Inference වෙන් කිරීම
- k-means, Hierarchical Clustering සහ DBSCAN basic idea එක තේරුම්ගැනීම
- PCA, Autoencoder awareness සහ anomaly score meaning තේරුම්ගැනීම
- Content-Based, Collaborative සහ Hybrid Recommendation වෙන් කිරීම
- Support, Confidence, Lift සහ Knowledge Graph basics තේරුම්ගැනීම
- Trend, Seasonality, Lag, baselines, ARIMA, Prophet සහ ML forecasting map එක දැනගැනීම
- Task-specific evaluation metrics සහ explainability limits තේරුම්ගැනීම

Today's 6 Video Map

1. Choosing the Correct Method Family
2. Clustering and Hidden Groups
3. Dimensionality Reduction and Anomaly Detection
4. Recommendations, Association Rules and Knowledge Graphs
5. Forecasting and Causal Questions
6. Evaluation, Explainability and Model Selection

Core rule: Problem → Required Output → Available Data → Task Type → Method Family → Validation → Metric → Risk → Human Decision.

Choosing the Correct Method Family

එකම problem wording එක වෙනස් වුණම method family එකක් වෙනස් වෙන්න පුළුවන්.

Classification

Known category එකක් predict කරනවා.

Spam / Not Spam, Disease A/B/C

Regression

Continuous numerical value එකක් predict කරනවා.

House price, income, temperature

Ranking

Items හොඳම එකෙන් පහළට order කරනවා.

Search results

Clustering

Labels නැති data වල similar groups හොයනවා.

Anomaly Detection

Unusual / rare cases හඳුනාගන්නවා.

Recommendation

Userට relevant items suggest හෝ rank කරනවා.

Association Rules

Items එකට පෙනෙන patterns හොයනවා.

Forecasting

Time order එක භාවිත කර future values estimate කරනවා.

Optimisation

Constraints ඇතුළත best action / allocation එක හොයනවා.

Causal Inference

Intervention එකක් නිසා outcome එක වෙනස් වුණද කියලා බලනවා.

Task, Method Family සහ Algorithm - එකම දෙයක් නොවේ

Classification = task type. **Supervised Learning** = broader learning approach. **Random Forest** = classification සඳහා use කරන්න පුළුවන් algorithm එකක්.

Problem → Output → Task Type → Method Family → Algorithm

Mini Task: Customer Segmentation, Suspicious Transaction, Movie Suggestion, Bread + Milk Pattern, Next Month Sales, Medical Category - මේ cases හයට method family එක match කරන්න.

Clustering and Hidden Groups

Clustering කියන්නේ labels නැති data වල similar observations groups වලට organize කරන Unsupervised Learning task එකක්.

k-means

k = clusters ගණන.

1. Initial centroids තෝරනවා
2. Points nearest centroid එකට assign කරනවා
3. Centroids update කරනවා
4. Stable වෙන තුරු repeat කරනවා.

Compact, roughly round clusters සඳහා useful. k කලින් තෝරන්න ඕන.

Hierarchical Clustering

මුලින් හැම point එකම වෙනම cluster එකක්.
ප්‍රශ්න clusters ටිකෙන් ටික merge කරනවා.

Dendrogram = merge hierarchy එක tree diagram එකකින් පෙන්වන visual එක.

DBSCAN

Density-based clustering.

eps = neighbourhood radius

min_samples = dense region එකක් ලෙස consider කරන්න අවශ්‍ය minimum nearby points.

Irregular shapes සහ noise points හඳුනාගන්න useful.

Scaling matters

Distance-based clustering වල feature scales වෙනස් නම් result distort වෙන්න පුළුවන්. Age 18-80 සහ Income 50,000-5,000,000 වගේ scales තිබුණොත් income numbers distance එක dominate කරන්න පුළුවන්.

How to Evaluate Clustering

- **Silhouette Score:** own cluster එකට fit වීම + other clusters වලින් separation
- **Stability:** sample එක ටිකක් වෙනස් කළත් similar clusters ලැබෙනවද?
- **Domain Usefulness:** business / real-world meaning තියෙනවද?

Important Caution

Cluster 0, 1, 2 කියන්නේ model identifiers විතරයි. "Cluster 0 = Good Customers" කියලා model එක automatically කියන්නේ නැහැ. Meaning එක human interpretation + domain knowledge මත තීරණය කරන්න ඕන.

Dimensionality Reduction and Anomaly Detection

Features ගොඩක් තියෙන data simplify කරන එක සහ unusual cases signal කරන එක.

PCA - Principal Component Analysis

PCA data එකේ වැඩිපුරම variation තියෙන directions හඳුනාගෙන fewer components වලින් represent කරන්න උදව් කරන method එකක්.

PC1 / PC2 original features නෙවේ. ඒවා original features කිහිපයක weighted combinations.

Explained Variance: selected components වලින් original variation කොච්චර capture කරනවද කියන indication එක.

High explained variance $\neq$ all important information preserved.

Autoencoder - Basic Awareness

Input $\rightarrow$ Encoder $\rightarrow$ Latent Representation $\rightarrow$ Decoder $\rightarrow$ Reconstruction

Neural network එක input එක compressed representation එකකට map කරලා නැවත reconstruct කරන්න ඉගෙනගන්නවා.

Dimensionality Reduction සහ Anomaly Detection සඳහා භාවිත වෙන්න පුළුවන්.

Anomaly Detection Methods

z-score / statistical methods

Isolation Forest

One-Class SVM

Autoencoder Reconstruction Error

Anomaly Score

Model එක observation එක කොච්චර unusual ද කියලා score එකක් දෙන්න පුළුවන්.

Threshold අඩු කළොත් alerts වැඩි, true anomalies වැඩිපුර catch වෙන්න පුළුවන්, නමුත් false alarmsත් වැඩි වෙන්න පුළුවන්.

Threshold වැඩි කළොත් alerts අඩු, නමුත් real anomalies miss වෙන්න පුළුවන්.

Very Important

Anomaly score fraud prove කරන්නේ නැහැ. ඒක investigation සඳහා signal එකක් විතරයි. Context, original features සහ human review අවශ්‍යයි.

Mini Task: 2D projection එකක far point එකක් දැක්කා කියලා ඒක automatic anomaly එකක්ද? - නැහැ. ඒ point එකේ original features සහ context check කරන්න.

Recommendations, Association Rules and Knowledge Graphs

Personalisation, co-occurrence patterns සහ connected relationships එකිනෙකට වෙනස් concepts.

Content-Based

User preference + item features compare කරනවා.

Advantage: new item features තිබුණොත් recommend කරන්න පුළුවන්.

Risk: userට කලින් කැමති දේවල් වගේම items නැවත නැවත පෙන්වන්න පුළුවන්.

Collaborative Filtering

Many usersගේ interactions භාවිත කරනවා.

User-Item Matrix: rows = users, columns = items, cells = rating/click/purchase/watch time.

Hybrid Recommendation

Content + collaborative + popularity + context + business rules combine කරන approach එක.

Cold Start: new user හෝ new item එකකට history අඩු problem එක.

Recommendation Quality ≠ Clicks Only

Relevance

Diversity

Coverage

Novelty

Fairness

Serendipity

User Control

Association Rules

Support: A සහ B දෙකම එකට පෙනෙන proportion.

Confidence: A තිබෙන cases අතරින් Bත් පෙනෙන proportion.

Lift: A තිබුණාම B පෙනෙන rate එක Bගේ normal rate එකට වඩා කොච්චර වෙනස්ද?

Association ≠ Causation.

Knowledge Graph

Entities / Nodes සහ Relationships / Edges graph structure එකක් ලෙස represent කරනවා.

Example: Course → teaches → Topic → requires → Skill.

Search, recommendation, fraud networks, question answering සහ reasoning support සඳහා useful.

Forecasting and Causal Questions

Forecasting future value estimate කරනවා. Causal inference intervention effect එක අහනවා.

Time Series Anatomy

long-term direction

repeating time pattern

variation irregular

information

old pattern changes

Forecast Baselines & Methods

- Naive Forecast
- Seasonal Naive
- Moving Average
- Exponential Smoothing
- ARIMA
- Prophet
- ML-based Forecasting with lag/calendar/external features

Time-Aware Validation

Forecasting වල random shuffle බොහෝවිට wrong. Future records training එකට leak වෙන්න පුළුවන්.

Past Training → Later Validation → Newest Test

Backtesting: historical points වල “එදා අපි හිටිය නම් next period කොහොම forecast කරයිද?” කියලා නැවත නැවත simulate කරන validation method එක.

Point Forecast vs Forecast Interval

Point Forecast: එක single estimate. Example: Next month sales = 12,000 units.

Forecast Interval: uncertainty පෙන්වන range. Example: 10,500-13,500. මේක 100% guarantee එකක් නොවේ.

Forecasting Question

“Next month sales කොච්චර වෙයිද?”



Causal Question

“10% discount දුන්නොත් sales කොච්චර වෙනස් වෙයිද?”

Reminder: chart එකක campaign එකෙන් පස්සේ sales වැඩි වුණ කියලා campaign එක cause එකයි කියලා alone prove කරන්න බැහැ. Seasonality, holidays, competitors, weather, economy වගේ confounders බලපාන්න පුළුවන්.

Evaluation, Explainability and Model Selection

Different tasks සඳහා different evidence අවශ්‍යයි. One metric fits all කියලා method එකක් නැහැ.

Classification

Accuracy, Precision, Recall, F1, ROC-AUC, Calibration.

Imbalanced data වල Accuracy alone misleading වෙන්න පුළුවන්.

Regression

MAE: average absolute error size.

RMSE: large errors වලට වැඩි penalty.

Forecasting

MAE/RMSE + baseline comparison + chronological backtesting + interval coverage.

Clustering

Silhouette + stability + domain usefulness.

Anomaly Detection

Useful alerts, missed anomalies, false alarm cost, investigation capacity.

Recommendation

Relevance, ranking quality, diversity, coverage, novelty, fairness, user control.

Robustness

Noise, missingness, subgroup changes, season changes හෝ distribution shift යටතේ performance stable ද කියලා stress-test කරනවා.

Explainability

Global explanation = overall model behaviour. Local explanation = one prediction ගැන.

Feature Importance / SHAP / LIME model behaviour explain කරන්න උදව් කරනවා. **Causality**, **correctness** හෝ **fairness** prove කරන්නේ නැහැ.

AutoML - Useful, but not the final decision maker

AutoML preprocessing, candidate model search සහ hyperparameter tuning automate කරන්න පුළුවන්. නැමැයි problem framing, legitimate target, privacy, fairness, causal validity, harm tolerance සහ accountability තීරණය කරන්න බැහැ.

AI PROBLEM SELECTION BOARD - WORKSHEET

Real-world problem එකක් analyse කරනකොට මේ board එක පුරවන්න. Method එක choose කරන්න කලින් problem structure එක පැහැදිලි කරන්න.

Case	Task / Method	Validation / Metric	Main Risk
Customer Segmentation	Clustering k-means / Hierarchical / DBSCAN	Silhouette + Stability + Domain Usefulness	Meaningless clusters
Suspicious Transaction	Anomaly Detection Isolation Forest	Useful alerts + missed anomalies + false alarms	Too many alerts / missed events
Movie Recommendation	Recommendation / Ranking	Relevance + Diversity + Coverage + User Value	Cold start / filter bubble
Market Basket	Association Rules	Support + Confidence + Lift	Association mistaken as causation
Monthly Sales	Forecasting	Chronological Backtesting + MAE / RMSE	Structural break / future leakage
Medical Category	Classification	Precision / Recall / F1 / ROC-AUC / Calibration	High-consequence errors

Now create your own case

Problem Question:
Required Output:
Available Data / Labels:
Task Type:
Method Family:
Validation Design:
Baseline:
Main Metric:
Additional Evaluation:
Main Risk / Harm:
Human Review Needed?

Final self-check: Method data structureට match ද? Metric real-world consequence එක represent කරනවද? Baseline තියෙනවද? Leakage, threshold, calibration, robustness, explainability සහ human review check කළාද?

DAY 4 QUICK SELF-CHECK

Quiz Questions

1. Method family තෝරන්න කලින් first question එක මොකක්ද?
2. Classification, Regression සහ Ranking අතර වෙනස?
3. k-means එකේ k represent කරන්නේ මොකක්ද?
4. Dendrogram පෙන්වන්නේ මොකක්ද?
5. DBSCAN noise mark කරන්නේ කොහොමද?
6. PCA component original feature එකක්ද?
7. Anomaly score fraud prove කරනවාද?
8. Content-Based vs Collaborative Recommendation?
9. Cold Start කියන්නේ මොකක්ද?
10. Support / Confidence / Lift meaning?
11. Forecasting වල random shuffle risky ඇයි?
12. Forecasting vs Causal Inference?
13. MAE vs RMSE?
14. Explainability causality prove කරනවාද?

Answer Hints

- Required output එක මොකක්ද කියලා මුලින් බලන්න.
- Category / number / ordered list.
- Chosen cluster count.
- Merge hierarchy.
- Low-density isolated points.
- නැහැ - weighted combination එකක්.
- නැහැ - investigation signal එකක්.
- Item/user features vs collective interactions.
- New user/item history නැති problem එක.
- Joint frequency / conditional frequency / relative lift.
- Future leakage වෙන්න පුළුවන්.
- Future estimate vs intervention effect.
- RMSE large errors වලට වැඩි penalty.
- නැහැ - model behaviour explain කරනවා.



Sampath Bandaranayake
Instructor - AI Panthiya

Need guided learning?

Online Zoom Classes - Contact Us on WhatsApp

WhatsApp: +81 70 2313 1981

MSc - AI and Digital Innovation (Final Year) - Tokyo International University, Japan

MBA - Cardiff Metropolitan University, UK

BSc - Applied Sciences, Sri Lanka

LinkedIn: <https://www.linkedin.com/in/sampathbandara/>

Website: <https://aipanthiya.com/>

Next step: **Day 5 - Deep Learning, Neural Networks, Transformers, Foundation Models and Generative AI**