

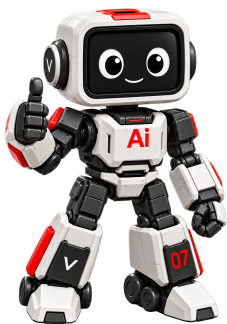


AI IN 7 DAYS

DAY 2 HANDBOOK - SELF-STUDY GUIDE

Data, Analytics, Mathematics and Statistics for AI - සරලව තේරුම්ගන්න

6 VIDEOS	60 MINUTES	1 DATASET	1 READINESS REPORT
-------------	---------------	--------------	-----------------------



© 2026 AI PANTHIYA

අද ඔබ ඉගෙනගන්නේ

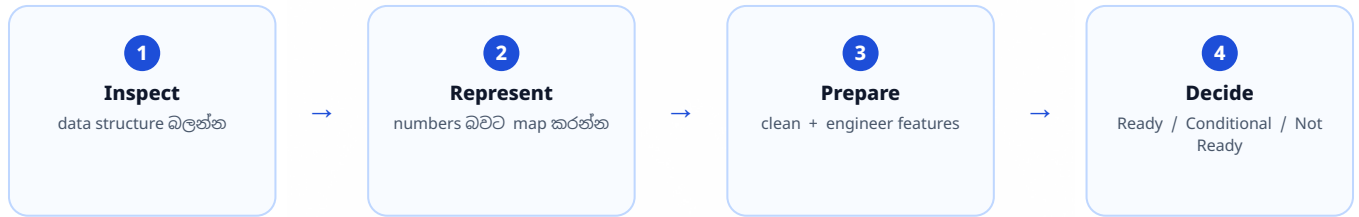
- Dataset structure සහ analytics question types
- Real-world data numbers බවට represent කරන ආකාරය
- Databases, ETL/ELT, batch, streaming සහ data pipelines
- AI සඳහා අවශ්‍ය math සහ statistics intuition
- EDA, cleaning, feature engineering සහ leakage
- Bias, causality, privacy සහ Data Readiness decision



Day 2 Main Task: Synthetic student performance dataset එකක් inspect කර **Data Readiness Report** එකක් සකස් කිරීම.

HOW TO USE THIS HANDBOOK

Day 2 එකේ formulas කටපාවිම් කිරීම වෙනුවට concept එක තේරුම්ගන්න, result එක interpret කරන්න, data problem හඳුනාගන්න සහ responsible decision එකක් ගන්න focus කරන්න.



Day 2 Learning Outcomes

- Observation, feature, identifier, target වෙන්කර හඳුනාගැනීම
- Descriptive, diagnostic, predictive, prescriptive සහ AI-assisted analytics වෙන්කර දැක්වීම
- Structured, semi-structured, unstructured data හඳුනාගැනීම
- Vectors, matrices, tensors සහ core statistics terms තේරුම්ගැනීම
- Missing values, duplicates, invalid values, outliers, imbalance, leakage හඳුනාගැනීම
- Bias, privacy, causality සහ readiness evaluate කිරීම

Today's 6 Video Map

1. Data, Analytics and the Questions We Ask
2. How Computers Represent the World
3. Databases and Data Engineering for Beginners
4. Mathematics and Statistics in One Connected Map
5. Exploratory Analysis, Cleaning and Feature Engineering
6. Bias, Causality, Privacy and Data Readiness

Scientific cautions: Outlier එකක් හැමවිටම error එකක් නෙවේ. Correlation එක causation prove කරන්නේ නැහැ. Scaling හැම algorithm එකකටම අනිවාර්ය නෙවේ. Clean dataset එකක් objective එකට unsuitable වෙන්නත් පුළුවන්.

Data, Analytics and the Questions We Ask

Dataset එක කියවන්න ඉගෙනගමු.

Dataset Structure

Observation

එක් row එකෙන් represent කරන unit එක. Example: student record.

Feature

analysis/model input ලෙස භාවිත කරන variable එක.

Identifier

record එක uniquely හඳුනාගන්න field එක. Example: student_id.

Target

model එක predict කරන්න ඉගෙනගන්න output variable එක.

Data Types

Numerical: attendance rate, score

Categorical: course language, internet access

Ordinal: low / medium / high

Date-time: recorded date

Free text: comments, reviews

Important: student_id identifier එකක්. Prediction feature එකක් ලෙස භාවිත කිරීම සාමාන්‍යයෙන් useful නෙවේ.

Analytics Question Ladder

Descriptive

මොකක්ද සිදු වුණේ?



Diagnostic

ඇයි සිදු වුණේ?



Predictive

ඉදිරියේ මොකක් වෙයිද?



Prescriptive

දැන් මොන action එක ගන්නද?



AI-assisted

patterns explain / summarize කිරීම

Mini Task: Student dataset එකක් ගෙන one observation, one identifier, three candidate features, one target සහ one predictive question ලියන්න.

How Computers Represent the World

Computer එකට photo, voice, sentence “directly” තේරෙන්නේ නැහැ. ඒවා numerical representations බවට පත් වෙනවා.

Image

Pixels → numbers. Colour image එකක සාමාන්‍යයෙන් Height × Width × RGB Channels.

1080 × 1920 × 3

Audio

Raw waveform හෝ spectrogram ලෙස represent කරන්න පුළුවන්.

Time → Amplitude

Categories

Nominal categories සඳහා one-hot encoding.
Ordinal categories සඳහා ordered encoding.

Yes/No → [1,0] / [0,1]

Text

Sentence → tokens → token IDs → embeddings.

“AI is useful” → tokens → vectors

Date & Time

year, month, weekday, hour, season වගේ useful features හදන්න පුළුවන්.

Sensor / Time Series

value + timestamp. Chronological order රැකගන්න ඕන.

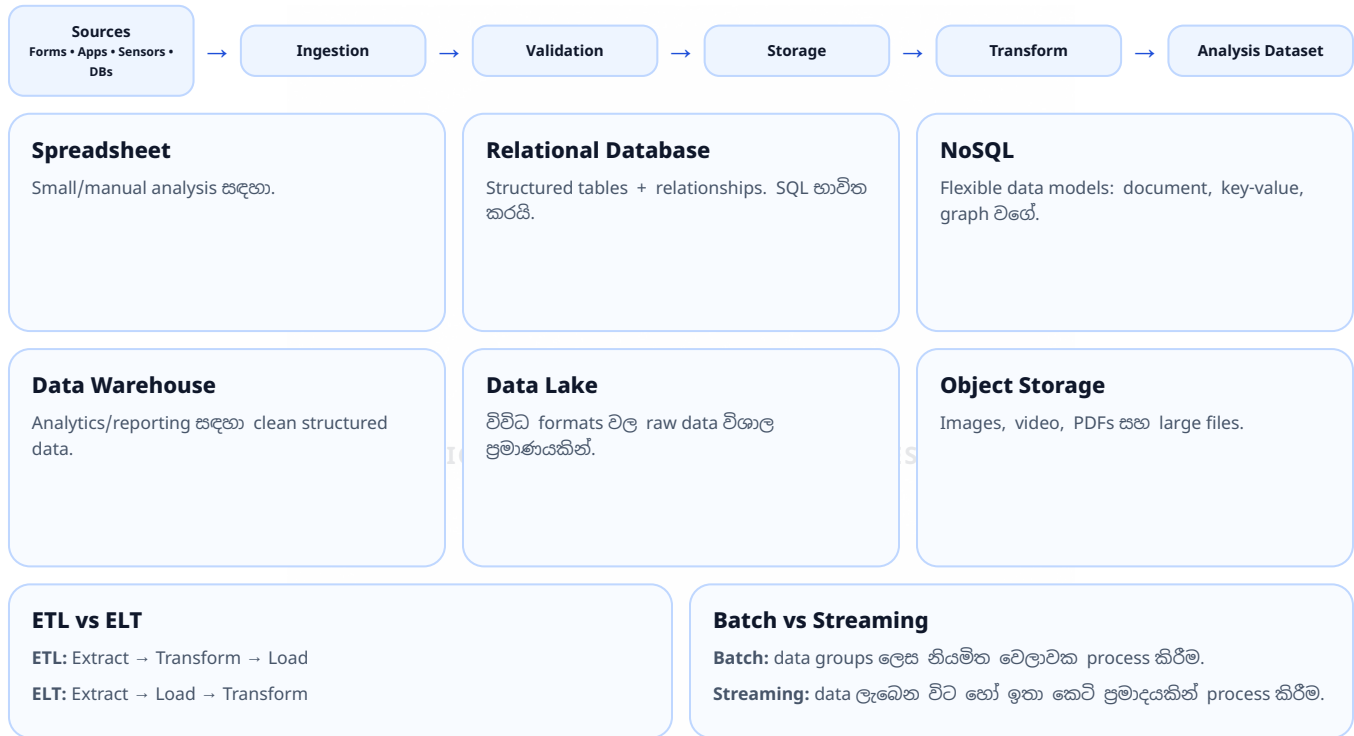
Scalar → Vector → Matrix → Tensor

Scalar → Vector → Matrix → Tensor
5 → [10,20,30] → rows × columns → dimensions 3+

Remember: Representation එක poor නම් powerful model එකක් තිබුණත් important information අහිමි වෙලා prediction quality අඩු වෙන්න පුළුවන්.

Databases and Data Engineering for Beginners

Raw data එක analysis-ready dataset එකක් වෙන්නේ කොහොමද?



Mini Task: Google Form + school database + attendance device → validation → warehouse → analysis dataset ලෙස ඔබේ own pipeline map එක අඳින්න.

Mathematics and Statistics in One Connected Map

Proofs කටපාවම් නොකර, AI තුළ math සහ statistics භාවිත වන “why” එක තේරුම්ගමු.

Math Side

- **Variable:** value එකක් represent කරන නම
- **Function:** input → output mapping
- **Vector:** ordered list of numbers
- **Matrix:** rows × columns
- **Tensor:** dimensions ඕනෑම ගණනක්
- **Distance:** points දෙක අතර සමානකම/දුර
- **Loss:** prediction error measure එක
- **Gradient:** loss වෙනස් වන direction
- **Optimisation:** loss අඩු කිරීමට parameters update කිරීම

Statistics Side

- **Mean:** average
- **Median:** middle value
- **Mode:** most frequent value
- **Range:** max - min
- **Variance / SD:** spread
- **Percentile:** relative position
- **Distribution:** values බෙදී ඇති overall pattern
- **Probability:** uncertainty measure
- **Correlation:** variables move together ද?

Model Learning - Core Loop

Prediction → Compare with Actual → Calculate Loss → Gradient
→ Update Parameters ↻

Important: Correlation එකක් ජ්‍යෙෂ්ඨ කියලා causal relationship එකක් automatically prove වෙන්නේ නැහැ.

Exploratory Analysis, Cleaning and Feature Engineering

Model එකට data දීමට පෙර dataset එක “තේරුම්ගෙන” සකස් කරන්න.

EDA - Exploratory Data Analysis

Shape & Columns → Data Types → Missing Values → Distributions
→ Categories → Relationships

Missing Values

Missingness reason එක තේරුම්ගන්න. Drop / impute blindly නොකරන්න.

Duplicates

True duplicate ද legitimate repeated event ද බලන්න.

Invalid Values

Example: attendance = 150%. Domain rules වලින් check කරන්න.

Outliers

Outlier = error කියලා assume නොකරන්න. Investigate කරන්න.

Encoding

Categorical data numbers බවට represent කරන්න.

Scaling

Features comparable scale එකකට ගෙන එන්න - method/task අනුව.

Feature Engineering

Existing data වලින් useful new features හදන්න.

Class Imbalance

One class too rare නම් evaluation සහ training strategy වෙනස් කරන්න වෙයි.

Data Leakage: prediction කරන සෑම වෙලාවේ ලබාගන්න බැරි information training data එකට accidentally ඇතුළත් වීම. Example: final exam score එක feature එකක් ලෙස දාලා final result කලින් predict කිරීම.

Bias, Causality, Privacy and Data Readiness

Dataset එක clean වුණට පමණක් enough නෙවේ. Objective එකට suitable ද? responsible ද?

Sampling / Selection Bias

Sample එක target population එක හෙඳින් represent නොකරන තත්ත්වය.

Measurement Bias

එකම thing groups අතර equally accurate ලෙස measure නෙවීම.

Label Bias

Target labels assign කරන ක්‍රමය subjective / inconsistent / unfair වීම.

Historical Bias

Past unfair patterns data තුළම තිබීම.

Prediction vs Causality

Prediction: “මේ student fail වෙන්න risk එක කොච්චරද?”

Causal question: “Extra class එකක් දුන්නොත් fail risk එක අඩු වෙයිද?”

Confounder: study hours සහ score දෙකටම motivation බලපාන්න පුළුවන්.

Privacy & Permission

Personal data minimise කරන්න. Access controls, permission, retention period, anonymisation/pseudonymisation ගැන සැලකිලිමත් වෙන්න.

READY

purpose clear, data suitable, major risks controlled

CONDITIONALLY READY

fixable issues තියෙනවා

NOT READY

major data, bias, privacy or leakage problem

Day 2 Main Task: Synthetic student dataset එකට Data Readiness Report එකක් පුරවලා final decision එක reason එකක් සමඟ දෙන්න.

DATA READINESS REPORT - WORKSHEET

පහත items පුරවලා Day 2 practical task එක complete කරන්න.

1. Intended problem

Dataset support කරන question / decision එක

2. Unit of observation

One row represent කරන්නේ කවුද?

3. Analytics question

Descriptive / Diagnostic / Predictive / Prescriptive

4. Data sources

Spreadsheet / DB / Form / Sensor / API

5. Candidate features

Input columns + reason

6. Candidate target

Target defined consistently ද?

7. EDA findings

missingness, distributions, categories, relationships

8. Quality problems

missing / duplicate / invalid / outlier

9. Representation needs

encoding / scaling / tokenization / date features

10. Feature engineering

create / select / remove features

11. Bias & coverage

underrepresented groups / measurement / labels

12. Causality limits

confounders? causal claim ද?

13. Leakage risks

future info / target proxy / preprocessing leakage

14. Privacy & permission

personal data / access / minimisation / retention

15. Final Decision

☐ Ready ☐ Conditionally Ready ☐ Not Ready

Reason:

Core idea: Clean data ≠ Suitable data. Dataset එක technically clean වුණත් target population එක, objective එක, time period එක හෝ ethical requirements support නොකරනවා නම් භාවිත කරන්න හොඳ නෑහැ.

DAY 2 QUICK SELF-CHECK

Quiz Questions

1. Feature එක සහ target එක අතර වෙනස මොකක්ද?
2. Structured vs unstructured data example දෙකක් දෙන්න.
3. Tokenization සහ embedding අතර වෙනස මොකක්ද?
4. ETL සහ ELT අතර මූලික වෙනස මොකක්ද?
5. Mean සහ median වෙනස් වන්නේ කොහොමද?
6. Outlier එකක් remove කිරීමට පෙර investigate කරන්න ඕන ඇයි?
7. Data leakage example එකක් දෙන්න.
8. Correlation causation prove නොකරන්නේ ඇයි?
9. Dataset එක "Conditionally Ready" වෙන්න පුළුවන් situation එකක් දෙන්න.

Answer Hints

- Feature = input; target = output/prediction goal.
- Structured = table; unstructured = image/text/audio/video.
- Tokenization text split කරයි; embedding learned vector representation එකක්.
- ETL: transform before load; ELT: load before transform.
- Mean outliersට sensitive; median middle value.
- Outlier error එකක්ද valid rare case එකක්ද බලන්න.
- Future info training feature එකකට accidentally ඇතුළත් වීම.
- Third variable/confounding තියෙන්න පුළුවන්.
- Fixable missingness, encoding, documentation, bias issue තිබේ නම්.



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 3 - How Machine Learning Actually Works**