EdSutra

Teaching guide

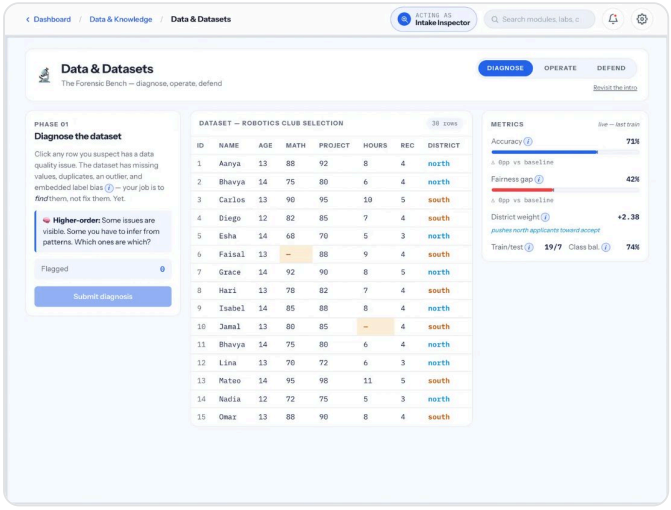
Module 04 · Data & Knowledge · Unit 01 · Stage 1, Foundations

Data & Datasets

The Data Kitchen

Grades 6–9 | **50 minutes** | Role: **Intake Inspector**

Grade 6: About age 11–12 · Year 7 in the UK · Class 6 in India



The big idea

A model is a mirror of its data, including the human choices baked into its labels. The robotics club's spreadsheet holds last year's committee decisions, which favored one district, so a model trained on it copies that pattern. Cleaning data is a set of judgment calls with trade-offs, and deleting the district column does not remove the bias.

The learner's job

Learners are the **Intake Inspector** at Greymill Depot · Goods In. This depot supplies every kitchen in the building. Nobody downstream can see what you let in. Their task: check the delivery before it becomes a model. Say who is missing.

By the end, learners can say

- I can read a dataset: rows are examples, columns are features, one column is the label
- I can identify problems in datasets: missing values, duplicates, outliers, label bias
- I can measure how cleaning choices change a real model's accuracy and fairness
- I can explain why deleting a biased feature doesn't remove bias

Before class

- One computer or tablet for each learner or pair, with EdSutra open (this Module is part of the full EdSutra studio).
- Open the studio once on each device before class, so the Unit still runs if the internet drops.
- Play the intro and the Diagnose tab yourself, so you know where learners will click.
- Write "Accuracy" and "Fairness gap" on the board. During Operate, collect each pair's best pipeline and both numbers.
- Best after What is AI? and How Machines Learn.
- Units open in order, and Module 1 comes first. To start a class here, switch on Open every Unit in Settings → Teachers & parents.

Words to introduce

Dataset: A table of examples: each row is one example, each column describes it.

Feature: One column the model can use, such as a math score.

Label: The answer column the model learns to copy; here, last year's decision.

Missing value: An empty cell where nobody wrote the number down.

Outlier: A value far from the rest, like an age of 99.

Train/test split: Rows the model learns from, and locked-away rows it is tested on.

Fairness gap: How differently the model accepts applicants from the two districts.

Running the lesson 50 minutes

Open 5 min

"A robotics club wants a computer to pick its new members because "it's objective." It will learn from last year's choices. Will it be fair?"

Before anyone clicks, ask: if last year's committee favored one district, would deleting the district column make the computer fair? Take a show of hands and come back to it in Apply.

Step	What learners do	Ask the class	Listen for
1. Remember 4 min	Read the case: the Northside Robotics Club wants a computer to pick members, trained on last year's 30 applications. Learners use the labeled sample table to name rows, features and the label.	• What does one row of the table stand for? • Which column is the label, and who wrote it?	Each row is one applicant, each column a feature, and the label is last year's human decision.
2. Understand 6 min	Learners answer two predictions to unlock the bench: what a model learns from people's past decisions, and why about 1 in 4 rows are locked away before training.	• Look at Omar and Esha. What pattern might the model copy? • Why test the model on rows it never studied?	The model copies the pattern in the labels, favoritism included. Only rows it never saw give an honest test.
3. Analyze 10 min	In Diagnose, learners click every row they suspect: missing values, duplicates, an outlier and label bias. They submit and see what they caught, wrongly flagged and missed.	• Which problems could you spot in a single row? • Which problems only showed up when you compared rows?	Gaps and an age of 99 show in one row. Copies and label bias only show when you compare rows.
4. Apply 15 min	In Operate, learners add cleaning steps, press Retrain and watch accuracy, the fairness gap and the district weight change. Four missions: lift accuracy, cut the gap, then both at once.	• What happened to the fairness gap when you dropped the district column? • Which steps helped accuracy, and which helped fairness? Were they the same? • Look back at our show of hands. Were we right?	Dropping district barely closes the gap, because the labels carry the bias. Accuracy and fairness can pull apart.
5. Evaluate 10 min	Defend needs Diagnose and two missions done. Learners name what each step in their pipeline costs, then decide whether to ship the model, with a reason true of their own numbers.	• Would you let this model pick next year's club? Why? • What can cleaning never fix in this dataset?	A call backed by the model's own numbers, which admits that cleaning cannot fix labels copied from biased decisions.

Steps follow Bloom's revised taxonomy, in the order the Unit presents them.

Deepen the thinking

Critical thinking	Two pipelines reach the same accuracy but different fairness gaps. How would you choose between them, and who should help decide?
AI thinking	If the club collected one new column next year, which would you add, and how might it quietly stand in for district?
Ethics and fairness	Reweighting boosts south-district accepts. When is it fair to adjust old decisions, and when could that create a new unfairness?
Connect to life	Think of a list or ranking in your life, like a team sheet or a top ten. Whose past choices shaped it?

Misconceptions to listen for

"The computer decided, so it's objective."
The labels are last year's human decisions. The district weight shows the model copied them: being from north pushes applicants toward accept.

"To remove bias, just delete the sensitive column."
In Operate the district column really disappears, yet the gap barely moves. The bias lives in the labels, and other columns can stand in for district.

"Cleaning data is just tidying, with one right answer."
Every step moves accuracy and fairness, sometimes in opposite directions. Mission 03 warns learners they will have to make a trade-off.

In real life the card learners see at the end of the Unit

- Some clubs, schools and companies sort applications with software. If it learned from past decisions, it copies their patterns.
- In a class survey spreadsheet, one copied row or an age typed as 99 skews the results.
- When news says "the computer decided", people still chose the data, the labels and what counted as fair.

Try it this week. Look at a table of results, like a sports league or class survey. Find one duplicate, missing or impossible value. What would a model learn from it?

Watch out for. Deleting a sensitive column doesn't remove bias. If past decisions were unfair, the bias lives in the labels and leaks through other clues.

Talk about it. Should a computer help choose who gets into a club or team? What would make that fair?

Exit ticket

1. In a dataset, what is a row, a feature and the label?

2. Why didn't deleting the district column remove the bias?

3. Name one cleaning step and what it costs.

Going further

The Data Explorer Lab's Spreadsheet station lets learners clean a messy sheet by dragging operations such as "Remove blank rows" onto it, then hunt for patterns in Explore. For homework, use the "Try it this week" activity from the In real life card.