Machine learning

IB DP Computer Science A4 Machine learning full package

A program that learns from examples

> Patterns from data, not rules from a programmer.

- Theme A4: what ML is, its types, how models are judged, and ethics.
- Prerequisites: scatter graphs and percentages.

AI, ML and deep learning

Deep learning is part of ML, which is part of AI.

- AI: computers doing tasks that seem to need human intelligence.
- Deep learning uses neural networks with many layers.

Three types of learning

Type	Learns from	Example
supervised	labelled data	spam or not; a price
unsupervised	unlabelled data	grouping shoppers
reinforcement	rewards and penalties	a robot learning to walk

Training and testing

> Test on data the model has never seen.

- **Overfitting**: memorises the training data, fails on new data.
- Underfitting: too simple to capture the pattern.

When 90% accuracy is useless

> Always saying "healthy" scores 90% and finds no one.

- **Accuracy** = correct / total x 100%: 90 of 100 is 90%.
- Check how the model does on each class.

Hardware for ML

Training

Thousands of cores in parallel.

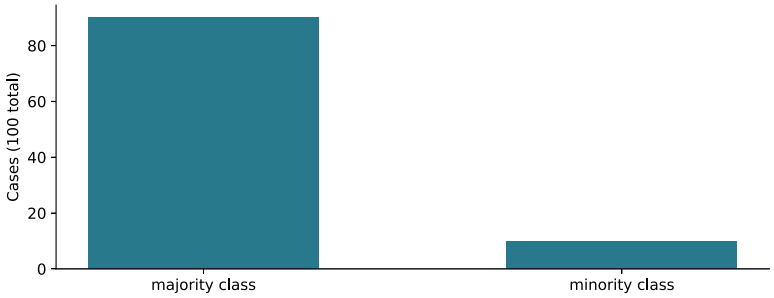
GPUs or special chips, often in the cloud.

Inference

Using the trained model.

Far less power - it can run on a phone.

Imbalanced classes

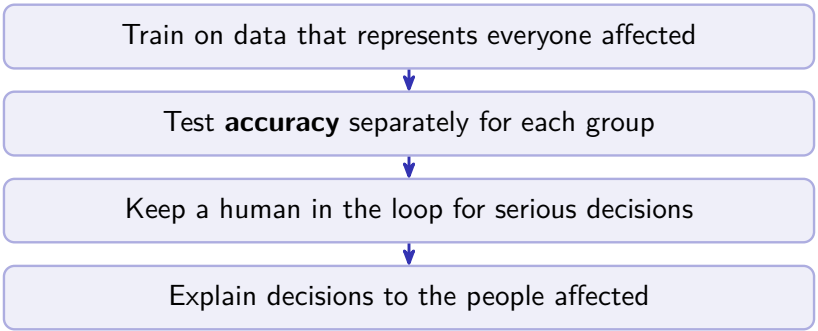


Accuracy alone can mislead

Who is affected?

Issue	What goes wrong
bias	face recognition less accurate for darker skin
transparency	a loan refused with no reason given
accountability	who answers for a self-driving crash?
privacy	photos scraped without consent

Reducing the harm



Preparing data

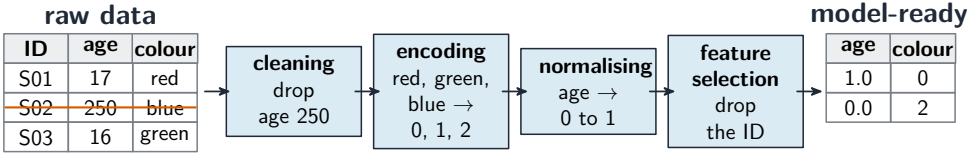
Step	**Example**
cleaning	delete a row with age 250
encoding	red, green, blue to 0, 1, 2
normalisation	income and age on a 0 to 1 scale
feature selection	drop the student ID number

Data leakage

> Excellent in testing, useless in real use.

- AHL: dimensionality reduction combines many features into fewer.
- Leakage: a **feature** not known when the prediction is made.

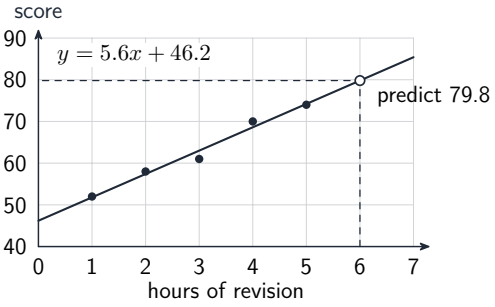
Preparing data for a model



Data leakage: never keep a feature that would not be known when the prediction is made

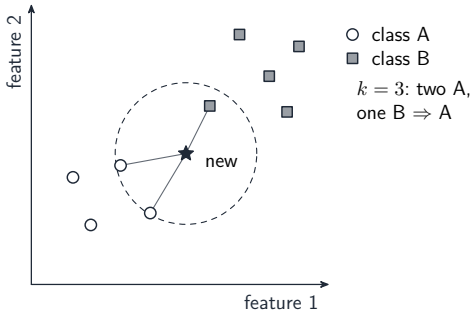
> Drop the ID, which carries no signal

Linear regression



$y = 5.6 \times 6 + 46.2 = 79.8$

k-nearest neighbours



k = 1 gives B; k = 3 gives A.

Decision trees and clustering

Decision tree

A sequence of yes/no questions about the features.

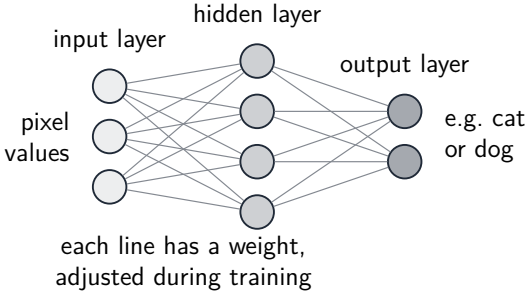
Each path ends at a leaf that gives the class.

k-means clustering

Choose k centres; join each item to the nearest.

Move each centre to its cluster's mean; repeat.

Neural networks



Training adjusts the weights to reduce the error.

Precision and recall

1. **Accuracy** (TP + TN) / total = (4 + 85) / 100 = 0.89
2. **Precision** TP / (TP + FP) = 4 / 14, about 0.29
3. **Recall** TP / (TP + FN) = 4 / 5 = 0.80

Decision

Recall: how many real positives found. Precision: how many flags were right.

Practice

Your turn

Try these:

A. Why test on data the model was not trained on?

B. 98% accuracy for a 2% disease: why may it be useless?

C. Predict the score for 3.5 hours with $y = 5.6x + 46.2$.

D. TP 30, FP 10, FN 20, TN 140: precision and recall?

Check yourself

