🧮 Module 1 – Descriptive Statistics and Distributions

Part A – Central Tendency and Spread

1. Marks: 50, 60, 65, 70, 75, 80, 85, 90
   - Find mean, median, mode.
   - Replace 90 with 900 and recompute mean. Explain the effect of outliers.

2. Temperatures (°C): 29, 31, 33, 33, 32, 31, 30
   - Calculate variance and standard deviation.
   - Which measure is easier to interpret and why?

Part B – Percentiles, IQR and Z-Score

3. Cat weights (kg): 2.5, 3.0, 3.2, 3.3, 3.4, 3.5, 3.6, 3.9, 4.0, 4.5
   - Find P25, P50, P75 and IQR. Mark possible outliers.

4. Given $\mu = 50$, $\sigma = 10$
   - Find Z for $x = 65$ and interpret.
   - For $Z = -2.5$, find x.

🎲 Module 2 – Probability Basics for ML

Part A – Basic Probability

1. In a survey of 100 people:
   - 40 like pizza (A)
   - 50 like burgers (B)
   - 20 like both ($A \cap B$)

   Find $P(A)$, $P(B)$, $P(A \cap B)$, and $P(A \mid B)$.

Part B – Conditional Probability and Bayes

2. Email dataset of 1 000 mails:

   - 100 spam (S)
   - 40 of these contain “free” (F)
   - 60 non-spam also contain “free”
     Compute $P(S \mid F)$ using Bayes’ Theorem →

     Explain what this means for a spam filter.

3. Disease example:

   - $P(\text{Disease}) = 1 / 10\,000$
   - $P(+ \mid D) = 0.99$
   - $P(- \mid \neg D) = 0.99$

     Find $P(D \mid +)$ and discuss the base-rate effect.

Part C – Confusion Matrix and Performance Metrics

Task 1:
A binary classifier results:

	Predicted Positive	**Predicted Negative**
Actual Positive	120	30
Actual Negative	80	770

Compute:

- Accuracy = (TP + TN) / Total
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 = 2 × Precision × Recall / (Precision + Recall)
- Specificity = TN / (TN + FP)
- NPV = TN / (TN + FN)
- Prevalence = (Actual Positive) / Total
  Explain what each metric tells you about model behavior.

Task 2:
A dataset of 10000 samples has only 2 % positives. The model predicts everything as negative.

- Write the confusion matrix.
- Compute Accuracy, Precision, Recall, F1, Specificity.
- Why is Accuracy misleading in this case?
- Which metric would be more appropriate for imbalanced data and why?

Hospital Triage Model [Optional]

A hospital triage model predicts the condition of patients as:
U = Urgent, **N = Normal**, and **R = Review (within 24 hours).**

On a test set of **1,200 patients**, the results are:

Actual \ Predicted	**U**	**N**	**R**
U	120	42	18
N	60	612	48
R	45	90	165

Tasks

1️⃣For each class (**U**, **N**, and **R**) — treat it as “positive” (one-vs-rest) — compute:

- Precision (PPV)
- Recall
- F1 Score
- NPV

2️⃣Compute **overall Accuracy** of the model.

3️⃣Combine **U** and **R** as one group of “Positive” (needs attention) and **N** as “Negative.”
Then compute:

- Accuracy
- Precision (PPV)
- Recall
- F1 Score
- NPV

4️⃣Briefly interpret which class or group performs best and why this matters in a medical triage context.