Partial Exam – I447 (E+F)/IN402 **2025 - 2026**

Duration: 1 hour **Total:** 30 points

Exercise 1: Core Concepts (6 pts)

$\frac{TP}{TP+FN}$

$\frac{TP}{TP+FP}$

1. Definitions (4 pts)

a. What is *overfitting* and how can it be detected?
b. Explain the difference between *bias* and *variance*.
c. What is the purpose of *regularization* in machine learning?
d. Which metric (Precision or Recall) is more important for Spam detection, and why?

2. True / False (2 pts)

Statement	T/F
Increasing polynomial degree increases variance.	
Logistic regression has a linear decision boundary in the feature space.	

Exercise 2 (5 pts): Given the dataset

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1

1. Compute the **information gain** for splitting on X_1 and on X_2.
2. Which feature will the tree choose first? Why?

Exercise 3 (8 pts): A binary classification model outputs probability

Sample	True y	Predicted P(y=1)
1	1	0.9
2	0	0.2
3	1	0.7
4	0	0.6

1. **(2 pt)** Compute the **binary cross-entropy loss** (average over all samples).
2. **(1 pt)** If threshold = 0.5, compute the predicted classes.
3. **(4 pts)** Build the **confusion matrix**, and compute: Accuracy, Precision, Recall and F1-score

4. **(1 pts)** Interpret what happens if we lower the threshold to 0.3 in terms of precision vs recall.

Exercise 4 (6 pts): Given the following true function

$$f(x) = \sin(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 0.1^2)$$

We focus at $x=\pi/2$, where: $f(\pi/2) = 1.0$. Irreducible error (variance of noise): $\sigma^2=0.01$
Predictions at $x=\pi/2$ across 5 datasets (D1–D5):

Model	Predictions	Expected Prediction	Bias²	Variance	Total Error
Linear	[0.9, 1.0, 0.8, 1.1, 0.95]	?	0.0025	0.012	0.0245
Polynomial	[1.05, 0.95, 1.15, 0.85, 1.1]	1	?	0.02	0.03
Neural Net	[1.02, 1.03, 0.98, 1.01, 0.99]	1.006	0.000036	?	?

a. For each of the three models, calculate the missing indicators
b. What is your conclusion about the results? Discuss and justify

Exercise 5 (5 pts): Given the following table

Threshold	TPR	FPR
0.9	0.2	0.01
0.7	0.5	0.05
0.5	0.8	0.20
0.3	0.95	0.50

a. Plot the ROC Curve
b. Which threshold minimizes false positives while keeping TPR $\geq$ 0.8?
c. If the cost of missing fraud (FN) is high, which threshold is best?