EXAM 1 STUDY GUIDE - WEEKS 1-4

Foundations of AI for Cybersecurity

CIS 350: AI for Cybersecurity - Smart Defense for the Digital Business

Colorado State University - Computer Information Systems

How to use this study guide

This guide pulls together everything Exam 1 covers (Weeks 1-4): security fundamentals; data, AI, and machine learning; supervised learning; and how we evaluate models. Each section gives the key ideas, the terms you must know, and self-check questions. Study actively - cover the answers and recall them from memory, then redo the worked confusion-matrix example. The exam is open-notes, individual work, with no heavy math and no coding.

Estimated review time: 45-60 minutes. Pair this with your lecture slides and Handout 4.

What Exam 1 Covers

Exam 1 is non-cumulative and covers Weeks 1-4 only. The four weeks build one pipeline: understand the security problem, gather and understand the data, build models from that data, and judge whether those models are any good. The exam rewards seeing how these connect, not memorizing isolated facts.

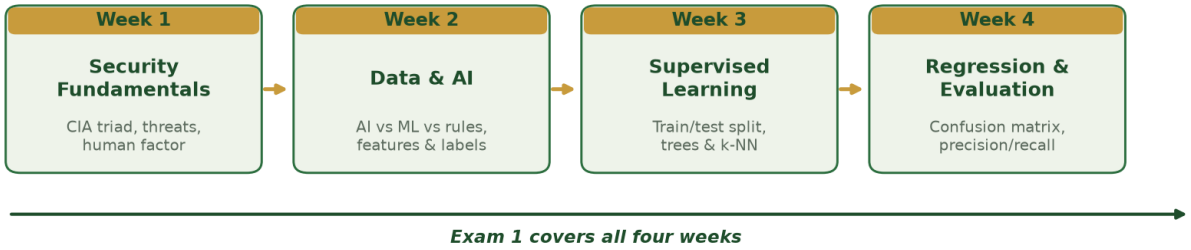

Figure 1. The four weeks form one pipeline: security context -> data -> models -> judging models.

Exam logistics

100 points, 90 minutes: multiple-choice, true/false, short answer, and applied scenarios. Online via Canvas, proctored with Respondus LockDown Browser. Open-notes and individual - no collaboration and no AI assistance. No heavy math and no coding. Install and TEST LockDown Browser before exam day.

Week 1 - Security Fundamentals

Core ideas

- CIA triad - the three goals of security: Confidentiality (keep data secret), Integrity (keep it accurate/unaltered), Availability (keep it usable when needed).
- Asset = what we protect; Threat = a potential source of harm; Vulnerability = a weakness that could be exploited; Risk = likelihood combined with impact.
- The human factor: people are the most targeted layer - phishing and social engineering exploit trust.
- Defense in layers: Prevention (block), Detection (notice), Response (contain and recover).

Watch out (common exam trap)

A THREAT is the potential harm or the attacker; a VULNERABILITY is the weakness they exploit. Don't swap them. Example: an unpatched server is a vulnerability; the hacker group that targets it is the threat.

Week 2 - Data, AI, and Machine Learning

Core ideas

- AI = machines doing tasks that seem to need human intelligence. ML = a subset of AI that LEARNS

patterns from data. Rule-based software follows fixed if/then rules a human wrote.
- Key difference: rules are written by a person; ML is learned from examples - so ML adapts as the data changes (e.g., spam that keeps evolving).
- A dataset has rows (examples), columns (features = input clues), and a label (the answer to predict).
- Security data sources: system logs, email, network traffic, user activity.
- 'Garbage in, garbage out': poor-quality data produces a poor model.

Week 3 - Supervised Learning

Core ideas

- Supervised learning = learning from LABELED examples (the correct answers are provided).
- Classification predicts a category (spam/ham); regression predicts a number (covered in Week 4).
- Train/test split: learn on the training set, then GRADE on a separate test set to estimate real skill.
- Overfitting = memorizing the training data and then failing on new data. Testing on training data hides this, which is why we hold out a test set.
- Two beginner classifiers: a decision tree (a learned flowchart of yes/no questions) and k-NN (classify by the majority label of the nearest neighbors; small k is noisy, large k is blurry).

> **The exam analogy for train/test**
>
> You don't grade students on the exact questions they studied - you test new ones. Same with models: we hold out a test set so the score reflects real skill, not memorization.

Week 4 - Regression and Model Evaluation

Core ideas

- Regression predicts a NUMBER (risk score, expected loss); classification predicts a category.
- Accuracy = percent correct - a good start, but it treats all mistakes as equal and can mislead when classes are imbalanced.
- Confusion matrix - four outcomes: True Positive, True Negative, False Positive (false alarm), False Negative (missed attack).
- Precision = of the items we flagged, how many were right (few false alarms). Recall = of the real positives, how many we caught (few misses). Improving one often lowers the other.
- The alert threshold is a dial you choose: lower it to catch more (higher recall, more false alarms); raise it for fewer false alarms (higher precision, more misses). Too many false alarms cause alert fatigue.

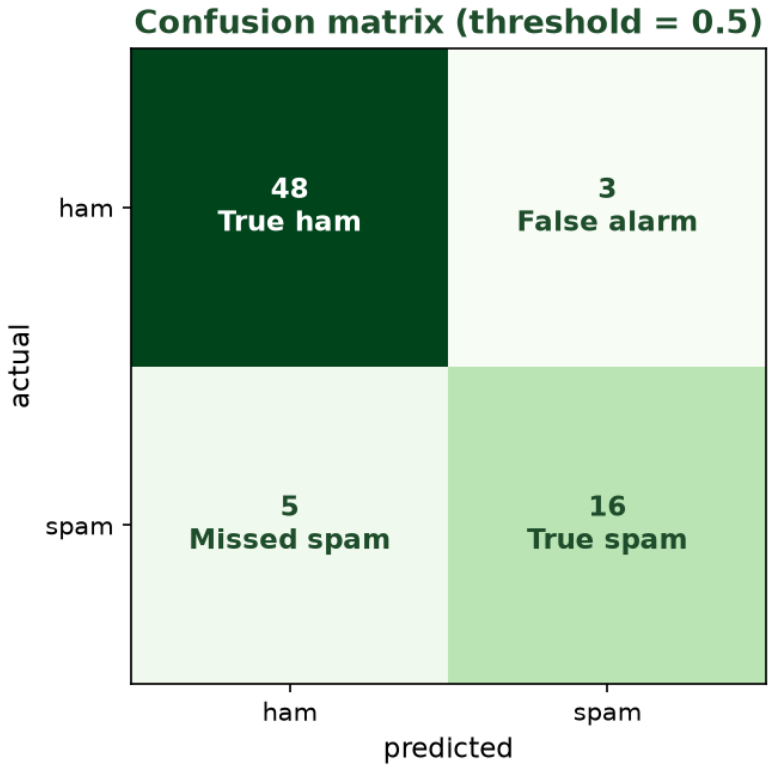

Figure 2. The confusion matrix: correct predictions on the diagonal, the two error types off it.

Worked Example: Reading a Confusion Matrix

A spam filter is tested on 100 emails and produces: 60 true negatives, 5 false positives, 10 false negatives, and 25 true positives. You should be able to answer all of the following:

- Accuracy = (correct) / (total) = (60 + 25) / 100 = 85%.
- False alarms (legitimate email wrongly flagged) = 5 (the false positives).
- Missed attacks (spam that slipped through) = 10 (the false negatives).
- For a hospital, LOWER the threshold to raise recall (miss fewer attacks), accepting more false alarms; reduce the downside with a quarantine folder, user training, and triage.

Key Terms to Know Cold

- CIA triad - Confidentiality, Integrity, Availability.
- Threat vs. Vulnerability vs. Risk.
- AI vs. Machine Learning vs. Rule-based software.
- Feature (input clue) and Label (the answer to predict).
- Training set / Test set; Overfitting.
- Decision tree; k-nearest neighbors (k-NN).
- Classification vs. Regression.
- Confusion matrix; True/False Positive; True/False Negative.
- Precision (few false alarms) vs. Recall (few misses).
- Threshold; alert fatigue.

Self-Check Questions

1. Name the three parts of the CIA triad and give an example of violating each.
2. Explain the difference between a threat and a vulnerability, with an example of each.
3. What makes machine learning different from rule-based software? Give a security example where ML wins.
4. Define a feature and a label using a phishing-email dataset.
5. Why do we split data into training and test sets? What goes wrong if we don't?
6. How does a decision tree make a prediction? How does k-NN?
7. Define false positive and false negative for a spam filter, and give the real-world cost of each.
8. Define precision and recall, and explain why they trade off.
9. Why can a 95%-accurate model still be a poor security tool?
10. For a hospital phishing filter, would you favor precision or recall, and why?

If you can answer all ten from memory and read a confusion matrix without hesitation, you are ready. Remember: AI assistance and collaboration are not permitted on the exam.

References

- NIST, Cybersecurity Framework 2.0, 2024.
- NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
- CISA, Phishing Guidance: Stopping the Attack Cycle at Phase One, 2023.
- OWASP, Machine Learning Security Top 10.
- Google, Machine Learning Crash Course (developers.google.com/machine-learning).
- scikit-learn User Guide: Model evaluation (metrics and scoring).