Q2 — True/False Questions (15 points)

Summary Table

#	Statement (abbreviated)	Answer	Key Reasoning
1	ABC frequent, BCD not → ABD cannot be frequent	**False**	BCD is not a subset of ABD; monotonicity doesn't apply sideways
2	More hash tables in LSH → fewer false positives	**False**	More bands increase candidate probability → more FPs (fewer FNs)
3	K-means handles outliers well via own clusters	**False**	K-means assigns outliers to nearest centroid; they distort clusters
4	Pruning combats decision tree overfitting	**True**	Pruning removes noisy branches, reducing model complexity
5	GAT learns a single attention weight per node	**False**	GAT learns attention weights per *edge* (neighbor pair), not per node
6	Cross-linking bloggers form a spider trap	**False**	Spider trap requires *all* out-links stay in group; blogs have external links too
7	Spam mass is always non-negative	**False**	Trusted pages can have TrustRank > PageRank → negative spam mass
8	NN embeddings are always linearly separable	**False**	No guarantee; depends on architecture, loss, and data
9	Root node always has highest information gain	**False**	Child nodes can have better splits relative to their local entropy
10	Greedy has higher competitive ratio than BALANCE	**False**	Greedy = 1/2; BALANCE = $1 - 1/e \approx 0.632$

Detailed Reasoning

1. False — Monotonicity applies to subsets, not siblings

The downward closure property says: if itemset I is frequent, every subset of I is also frequent. ABD's subsets are {AB, AD, BD, A, B, D} — none of which is BCD. The frequency of BCD is irrelevant to ABD.

2. False — More hash tables increase false positives

With b bands and r rows per band, candidate probability = 1 − (1 − s^r)^b. Increasing b (more hash tables) raises this probability for *all* pairs, including dissimilar ones → more false positives. (But fewer false negatives.)

3. False — K-means is sensitive to outliers

K-means minimizes within-cluster sum of squares with fixed K. Outliers pull centroids toward them rather than being isolated. This distorts the resulting clusters.

4. True — Pruning is a standard regularization method

Post-pruning removes subtrees that don't improve validation performance. Pre-pruning stops splits early. Both reduce overfitting.

5. False — Attention is per-edge

GAT computes: $e_{vu} = a(W \cdot h_v, W \cdot h_u)$, then normalizes via softmax: $\alpha_{vu} = \text{softmax}_u(e_{vu})$. Each neighbor u of v receives a *different* attention weight.

6. False — Spider trap requires no escape links

A spider trap is a set of nodes with no outgoing edges to nodes outside the set. Bloggers adding mutual links doesn't remove their existing external links. Also, a "loop" alone is not what defines a spider trap.

7. False — Spam mass can be negative for trusted pages

$\text{spam\_mass}(p) = (r_p - r_p^+) / r_p$. TrustRank teleports only to trusted pages, so trusted pages can receive more TrustRank than PageRank, yielding $r_p^+ > r_p$ and negative spam mass.

8. False — No such guarantee

Linear separability depends on the specific model, objective, and data. For example, a shallow autoencoder may produce embeddings that are *not* linearly separable.

9. False — Information gain depends on local data distribution

The root operates on the full dataset, but a child node may have a near-perfect split available. Example: root IG = 0.3 on a mixed dataset, but a child with balanced binary labels gets IG = 1.0 from a feature that perfectly separates its subset.

10. False — BALANCE beats Greedy

- Greedy competitive ratio: **1/2**
- Generalized BALANCE competitive ratio: **$1 - 1/e \approx 0.632$**

BALANCE achieves a strictly higher competitive ratio.