Artificial Intelligence

Decision Tree Design with ID3

Part A: Scenario and Features

This classification task considers a common decision faced by international students: whether they return to their home country during the winter holiday period. Each data instance represents an international student studying at a university abroad. The input features capture academic, financial, and administrative factors that typically influence this decision, while the class label represents the final travel outcome.

The aim of the classifier is to identify decision rules that describe how these factors combine to affect the student's choice. The target variable is binary: *Return Home = Yes* if the student travels back, and *No* otherwise.

The dataset includes four categorical input features. **Financial Situation** describes the student's financial circumstances and takes *Tight*, *Stable*, or *Comfortable*. **Exam Schedule Proximity** indicates how close assessment deadlines or examinations are to the holiday period, but does not necessarily prevent travel (e.g online exams); it can take the values *Close*, *Moderate*, or *Far*. **Length of Holiday Break** represents the duration of the winter break and is categorised as *Short*, *Medium*, or *Long*. **Visa Flexibility** reflects how easily the student can travel and re-enter the host country, considering factors such as visa conditions and re-entry requirements, and takes the values *Difficult*, *Moderate*, or *Easy*.

Features:

- Financial Situation ∈ {Tight, Stable, Comfortable}
- Exam Schedule Proximity ∈ {Close, Moderate, Far}
- Length of Holiday Break ∈ {Short, Medium, Long}
- Visa Flexibility ∈ {Difficult, Moderate, Easy}
- Class Label: Return Home ∈ {Yes, No}

#	Financial Situation	Exam Schedule Proximity	Length of Holiday Break	Visa Flexibility	Return Home
1	Comfortable	Far	Long	Easy	Yes
2	Comfortable	Moderate	Long	Moderate	Yes
3	Comfortable	Close	Medium	Easy	Yes
4	Comfortable	Close	Short	Moderate	No
5	Comfortable	Far	Medium	Difficult	No

6	Comfortable	Moderate	Medium	Difficult	No
7	Stable	Far	Long	Easy	Yes
8	Stable	Moderate	Medium	Moderate	Yes
9	Stable	Close	Short	Easy	No
10	Stable	Close	Medium	Difficult	No
11	Stable	Far	Short	Moderate	Yes
12	Stable	Moderate	Long	Difficult	No
13	Stable	Far	Medium	Easy	Yes
14	Tight	Far	Long	Easy	Yes
15	Tight	Close	Short	Difficult	No
16	Tight	Moderate	Medium	Easy	Yes
17	Tight	Close	Medium	Moderate	No
18	Tight	Far	Medium	Moderate	Yes
19	Tight	Close	Long	Easy	Yes
20	Tight	Moderate	Short	Moderate	No

Part B: Entropy and Root Information Gain

Let S denote the full training dataset consisting of 20 examples.
Class Entropy H(S):
In the full dataset:

- Number of examples with *Return Home = Yes*: 11
- Number of examples with *Return Home = No*: 9

The corresponding probabilities are:

$$p(Yes) = 11/20 = 0.55, \quad p(No) = 9/20 = 0.45$$

Using base-2 logarithms, the class entropy is computed as:

$$H(S) = -p(Yes)log2\,p(Yes) - p(No)log2\,p(No)$$

$$H(S) = -(0.55)\,log2(0.55) - (0.45)log2\,p(0.45) = 0.9928$$

Thus, the entropy of the full training set is: $H(S) = 0.9928$

Information Gain Analysis:
For each input feature, the conditional entropy after splitting and the corresponding information gain are computed using:

$$IG(S,\ A)\ =\ H(S)\ -\ \sum_{v\ \in\ Values(A)}\ \frac{|Sv|}{|S|}H(Sv)$$

Financial Situation: The dataset is split into three subsets: *Tight*, *Stable*, and *Comfortable*.

- Tight: |S|=7, Yes=4, No=3, H=0.9852
- Stable: |S|=7, Yes=4, No=3, H=0.9852
- Comfortable: |S|=6, Yes=3, No=3, H=1.0000

The conditional entropy is:

$$H(S\ |\ Financial)\ =\ 7/20\ (0.9852)\ +\ 7/20(0.9852)\ +\ 6/20(1.0000)\ =\ 0.9897$$

The information gain is:

$$IG(S,\ Financial)\ =\ 0.9928\ -\ 0.9897\ =\ 0.0031$$

Exam Schedule Proximity: The dataset is split into *Close*, *Moderate*, and *Far*.

- Close: |S|=7, Yes=2, No=5, H=0.8631
- Moderate: |S|=6, Yes=3, No=3, H=1.0000
- Far: |S|=7, Yes=6, No=1, H=0.5917

The conditional entropy is:

$$H(S\ |\ Exam)\ =\ 7/20\ (0.8631)\ +\ 6/20(1.0000)\ +\ 7/20(0.5917)\ =\ 0.8092$$

The information gain is:

$$IG(S,\ Exam)\ =\ 0.9928\ -\ 0.8092\ =\ 0.1836$$

Length of Holiday Break: The dataset is split into *Short*, *Medium*, and *Long*.

- Short: |S|=5, Yes=1, No=4, H=0.7219
- Medium: |S|=9, Yes=5, No=4, H=0.9911
- Long: |S|=6, Yes=5, No=1, H=0.6500

The conditional entropy is:

$$H(S\ |\ Break)\ =\ 5/20\ (0.7219)\ +\ 9/20(0.9911)\ +\ 6/20(0.6500)\ =\ 0.8215$$

The information gain is:

$$IG(S,\ Break)\ =\ 0.9928\ -\ 0.8215\ =\ 0.1713$$

Visa Flexibility: The dataset is split into *Difficult*, *Moderate*, and *Easy*.

- Difficult: |S|=5, Yes=0, No=5, H=0

- Moderate: |S|=7, Yes=4, No=3, H=0.9852
- Easy: |S|=8, Yes=7, No=1, H=0.5436

The conditional entropy is:

$$H(S \mid Visa) = 5/20\,(0) + 7/20(0.9852) + 8/20(0.5436) = 0.5623$$

The information gain is:

$$IG(S, Visa) = 0.9928 - 0.5623 = 0.4305$$

Root Node Selection

The information gain values for all features:

- Financial Situation: IG = 0.031
- Exam Schedule Proximity: IG = 0.1836
- Length of Holiday Break: IG = 0.1713
- Visa Flexibility: IG = 0.4305

The highest information gain is achieved by Visa Flexibility, therefore ID3 selects it as the root node.

Part C: Recursive Splitting (ID3)

From Part B, the highest information gain is obtained by **Visa Flexibility**, so we split *S* into three subsets:

$$S = S_{Difficult} \cup S_{Moderate} \cup S_{Easy}$$

- $S_{Difficult}$: |S| = 5, Yes - 0, No -5
- $S_{Moderate}$: |S| = 7, Yes - 4, No - 3
- S_{Easy}: |S| = 8, Yes - 7, No - 1

Branch 1: $S_{Difficult}$ (Visa = Difficult)

Remaining examples in subset

#	Financial Situation	Exam Schedule Proximity	Length of Holiday Break	Return Home
5	Comfortable	Far	Medium	No
6	Comfortable	Moderate	Medium	No
10	Stable	Close	Medium	No
12	Stable	Moderate	Long	No
15	Tight	Close	Short	No

Entropy:

$$p(Yes) = 0, \quad p(No) = 1$$

$$H(Sdifficult) = -(0)\,log2(0) - (1)log2\,p(1) = 0$$

Leaf Node: Return Home = No

Branch 2: S_{Easy} (Visa = Easy)

Remaining examples in subset

#	Financial Situation	Exam Schedule Proximity	Length of Holiday Break	Return Home
14	Tight	Far	Long	Yes
16	Tight	Moderate	Medium	Yes
7	Stable	Far	Long	Yes
9	Stable	Close	Short	No
1	Comfortable	Far	Long	Yes
3	Comfortable	Close	Medium	Yes
19	Tight	Close	Long	Yes
13	Stable	Far	Medium	Yes

Entropy of the subset:

$$H(Seasy) = -(7/8)\,log2(7/8) - (1/8)log2\,p(1/8) = 0.5436$$

Now compute information gain for remaining features: Financial Situation, Exam Schedule Proximity, Length of Holiday Break.

1. Split by Financial Situation (within S_{Easy})

- Tight: 3 (3 Yes, 0 No), H = 0
- Stable: 3 (2 Yes, 1 No), H = 0.9183
- Comfortable: 2 (2 Yes, 0 No), H = 0

Conditional entropy:

$$H(Seasy \mid Financial) = 3/8\,(0) + 3/8(0.9183) + 2/8(0) = 0.3444$$

Information gain:

$$IG(Seasy,\ Financial) = 0.5436 - 0.3444 = 0.1992$$

2. Split by Exam Schedule Proximity (within S_{Easy})

- Close: 3 (2 Yes, 1 No), H = 0.9183
- Moderate: 1 (1 Yes, 0 No), H = 0
- Far: 4 (4 Yes, 0 No), H = 0

Conditional entropy:

$$H(Seasy \mid Exam) = 3/8\,(0.9183) + 1/8(0) + 4/8(0) = 0.3444$$

Information gain:

$$IG(Seasy, Exam) = 0.5436 - 0.3444 = 0.1992$$

3. Split by Length of Holiday Break (within S_{Easy})

- Short: 1 (0 Yes, 1 No), H = 0
- Medium: 3 (3 Yes, 0 No), H = 0
- Long: 4 (4 Yes, 0 No), H = 0

Conditional entropy:

$$H(Seasy \mid Break) = 1/8\,(0) + 3/8(0) + 4/8(0) = 0$$

Information gain:

$$IG(Seasy, Break) = 0.5436 - 0 = 0.5436$$ ***(Highest IG)***

Highest IG is obtained by **Length of Holiday Break**. Leaves after splitting S_{Easy} by Break:

- Break = Short → (0 Yes, 1 No) → **Leaf No**
- Break = Medium → (3 Yes, 0 No) → **Leaf Yes**
- Break = Long → (4 Yes, 0 No) → **Leaf Yes**

Branch 3: $S_{moderate}$ (Visa = Moderate)

Remaining examples in subset

#	Financial Situation	Exam Schedule Proximity	Length of Holiday Break	Return Home
17	Tight	Close	Medium	No
8	Stable	Moderate	Medium	Yes
11	Stable	Far	Short	Yes
2	Comfortable	Moderate	Long	No
4	Comfortable	Close	Short	No
18	Tight	Far	Medium	Yes
16	Tight	Moderate	Medium	No

Entropy of the subset:

$$H(Smoderate) = -(4/7)\,log2(4/7) - (3/7)log2\,p(3/7) = 0.9852$$

Compute information gain for remaining features: Financial Situation, Exam Schedule Proximity, Length of Holiday Break.

1. Split by Financial Situation (within $S_{moderate}$)

- Tight: 3 (1 Yes, 2 No), H = 0.9183
- Stable: 2 (2 Yes, 0 No), H = 0
- Comfortable: 2 (1 Yes, 1 No), H = 1.0000

Conditional entropy:

$$H(Smoderate|\,Financial) = 3/7\,(0.9183) + 2/7(0) + 2/7(1.0000) = 0.6793$$

Information gain:

$$IG(Smoderate,\,Financial) = 0.9852 - 0.6793 = 0.3059$$

2. Split by Exam Schedule Proximity (within $S_{moderate}$)

- Close: 2 (0 Yes, 2 No), H = 0
- Moderate: 3 (2 Yes, 1 No), H = 0.9183
- Far: 2 (2 Yes, 0 No), H = 0

Conditional entropy:

$$H(Smoderate\,|\,Exam) = 2/7\,(0) + 3/7(0.9183) + 2/7(0) = 0.3936$$

Information gain:

$$IG(Seasy,\,Exam) = 0.9852 - 0.3936 = 0.5916$$ ***(Highest IG)***

3. Split by Length of Holiday Break (within $S_{moderate}$)

- Short: 3 (1 Yes, 2 No), H = 0.9183
- Medium: 3 (2 Yes, 1 No), H = 0.9183
- Long: 1 (1 Yes, 0 No), H = 0

Conditional entropy:

$$H(Smoderate\,|\,Break) = 3/7\,(0.9183) + 3/7(0.9183) + 1/7(0) = 0.7869$$

Information gain:

$$IG(Smoderate,\,Break) = 0.9852 - 0.7869 = 0.1983$$

Highest IG is obtained by **Exam Schedule Proximity.** Sub-branches after splitting $S_{Moderate}$ by Exam:

- Exam = Close → H = 0 → **Leaf No**
- Exam = Far → H = 0 → **Leaf Yes**

- Exam = Moderate:

Subset $S_{Moderate,\ Exam=Moderate}$ contains:

#	Financial Situation	Length of Holiday Break	Return Home
8	Stable	Medium	Yes
2	Comfortable	Long	Yes
20	Tight	Short	No

Entropy:

$$H = \ -\ (2/3)\ log2(2/3)\ -\ (1/3)log2\ p(1/3)\ =\ 0.9183$$

Remaining features: **Financial Situation** and **Length of Holiday Break.**

1. Split by Financial Situation ($S_{Moderate,\ Exam=Moderate}$)

- Tight: No (entropy = 0)
- Stable: Yes (entropy = 0)
- Comfortable: Yes (entropy = 0)

Conditional entropy:

$$H(\cdot\ |\ Financial)\ =\ 0$$

Information gain:

$$IG\ =\ 0.9183\ -\ 0\ =\ 0.9183$$

2. Split by Length of Holiday Break ($S_{Moderate,\ Exam=Moderate}$)

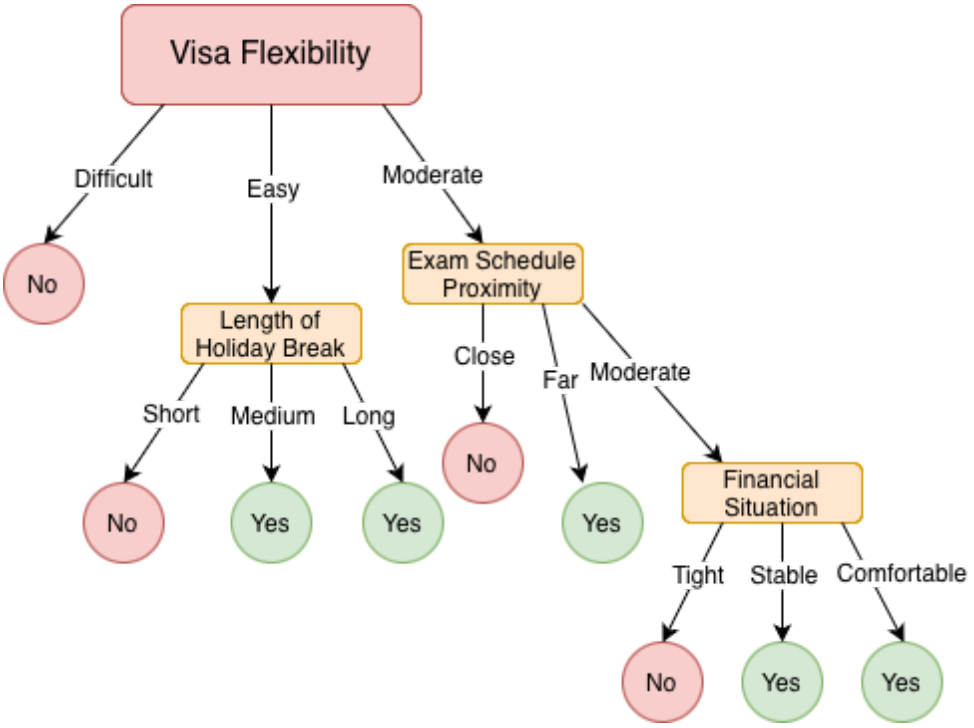
- Medium: Yes (entropy = 0)
- Long: Yes (entropy = 0)
- Short: No (entropy = 0)

So similarly:

$$IG\ =\ 0.9183\ -\ 0\ =\ 0.9183$$

Both features perfectly separate this subset. We select Split Node: Financial Situation.

Part D: Coloured Decision Tree Diagram



Decision nodes are shown as rectangular boxes and represent categorical input features selected by the ID3 algorithm based on information gain. The root node, *Visa Flexibility*, is highlighted at the top of the tree, as it provides the highest information gain. Internal decision nodes are coloured in light orange to distinguish them from leaf nodes. Leaf nodes are represented as circles and indicate the predicted class label: green circles correspond to *Return Home = Yes*, while red circles correspond to *Return Home = No*. Edges between nodes are labelled with the specific feature values that define each branch. The tree is constructed recursively until all branches terminate in pure leaf nodes, illustrating how administrative, academic, and financial factors combine to produce an interpretable classification outcome.

Part E: Narrative Explanation and Classification of Test Examples
(*Written with the assistance of ChatGPT (GPT-5.2, January 2026)*)

The decision tree predicts whether an international student returns home during the winter holidays by following a sequence of feature-based rules. The root node is Visa Flexibility. If visa flexibility is Difficult, the tree immediately predicts Return Home = No, reflecting that restrictive re-entry conditions strongly discourage travel. If visa flexibility is Easy, the tree next checks Length of Holiday Break: a Short break leads to No, while Medium or Long breaks lead to Yes, since sufficient time makes travel more worthwhile. If visa flexibility is Moderate, the tree considers Exam Schedule Proximity: Close exams predict No, Far exams predict Yes, and Moderate exam proximity triggers a final check of

Financial Situation, where Tight predicts No, while Stable or Comfortable predicts Yes.

Test instance 1

Financial = Comfortable, Exam = Close, Break = Long, Visa = Moderate
Path: Visa=Moderate → Exam=Close → Leaf: No.
Prediction: Return Home = No.
Comment: This seems plausible because exams close to the break may prevent travel even when finances and holiday length are favourable.

Test instance 2

Financial = Tight, Exam = Far, Break = Medium, Visa = Easy
Path: Visa=Easy → Break=Medium → Leaf: Yes.
Prediction: Return Home = Yes.
Comment: This is reasonable because an easy visa and a non-short break enable travel, although in reality tight finances could still prevent it (a limitation of this branch).

Limitations

1. The dataset is small and hand-crafted, so the tree may overfit and not generalise well to other student populations.
2. The tree uses coarse categories and cannot model trade-offs smoothly (e.g., very tight finances might override other factors), since it outputs hard rules rather than probabilities.

CIFAR-100 Semantic Expansion *(explanation was written with the help of ChatGPT (GPT-5.2, January 2026) to improve clarity and wording)*

Methodology & Rationale

The final embedding model was developed using a two-stage approach combining Labs 6 and 7. A Skip-Gram model was first trained on the Visual Genome corpus to obtain a stable base embedding space. To incorporate the full set of CIFAR-100 classes, two options were considered: retraining the model from scratch with an expanded vocabulary or extending the existing embedding space. Retraining was not chosen because CIFAR-100 class names lack a natural textual corpus and full retraining risked degrading the semantic structure already learned from Visual Genome. Instead, the evolutionary insertion approach was selected. This method allows missing CIFAR-100 tokens to be inserted individually using anchor words from the existing vocabulary, ensuring semantically plausible placement while preserving the original embeddings. As a result, the embedding space can be expanded in a controlled and interpretable manner without destabilising the base model.

Training Evidence

The base Skip-Gram model was refined through controlled hyperparameter experiments, evaluated using nearest-neighbor inspection and cosine similarity statistics. Negative sampling was varied from 8 to 15. While 8 negatives produced noisy neighborhoods and 15 led to overly uniform similarities, a value of 10 provided the best balance between stability and semantic coherence. Context size was also explored. A large context size of 4 captured broader graph structure but resulted in diffuse and less interpretable neighborhoods, while a context size of 1 improved locality but missed higher-level relations. A compromise value of 2 was therefore selected. An embedding dimension of 64 was used in the final model. Larger dimensions were considered but could not be reliably trained due to hardware constraints, making 64 a stable and practical choice. Several vocabulary normalization strategies, including stop-word and punctuation removal, were tested; however, they did not consistently improve semantic quality and were therefore not applied. To further stabilise training, the negative sampling distribution was adjusted to follow a unigram distribution raised to the power of 0.75, using node degree as a proxy for token frequency. This reduced the influence of high-degree nodes and improved training stability. For CIFAR-100 expansion, 32 class names were already present in the Visual Genome vocabulary. The remaining 68 classes were inserted using the evolutionary strategy, achieving an average cosine similarity of approximately 0.80 with their anchor sets.

Validation and Evaluation

The final model was validated using qualitative semantic analysis. Nearest-neighbor inspection confirmed that both original and inserted words occupied semantically plausible regions of the embedding space. For example, *beaver* clustered near *water* and *animal*. Anchor similarity analysis verified that inserted words remained close to their guiding concepts without collapsing the embedding space. Random cosine similarity statistics confirmed that sufficient variance was preserved. Across all tested configurations, the selected model consistently produced the most coherent neighborhoods and stable behavior and was therefore chosen as the final submission.