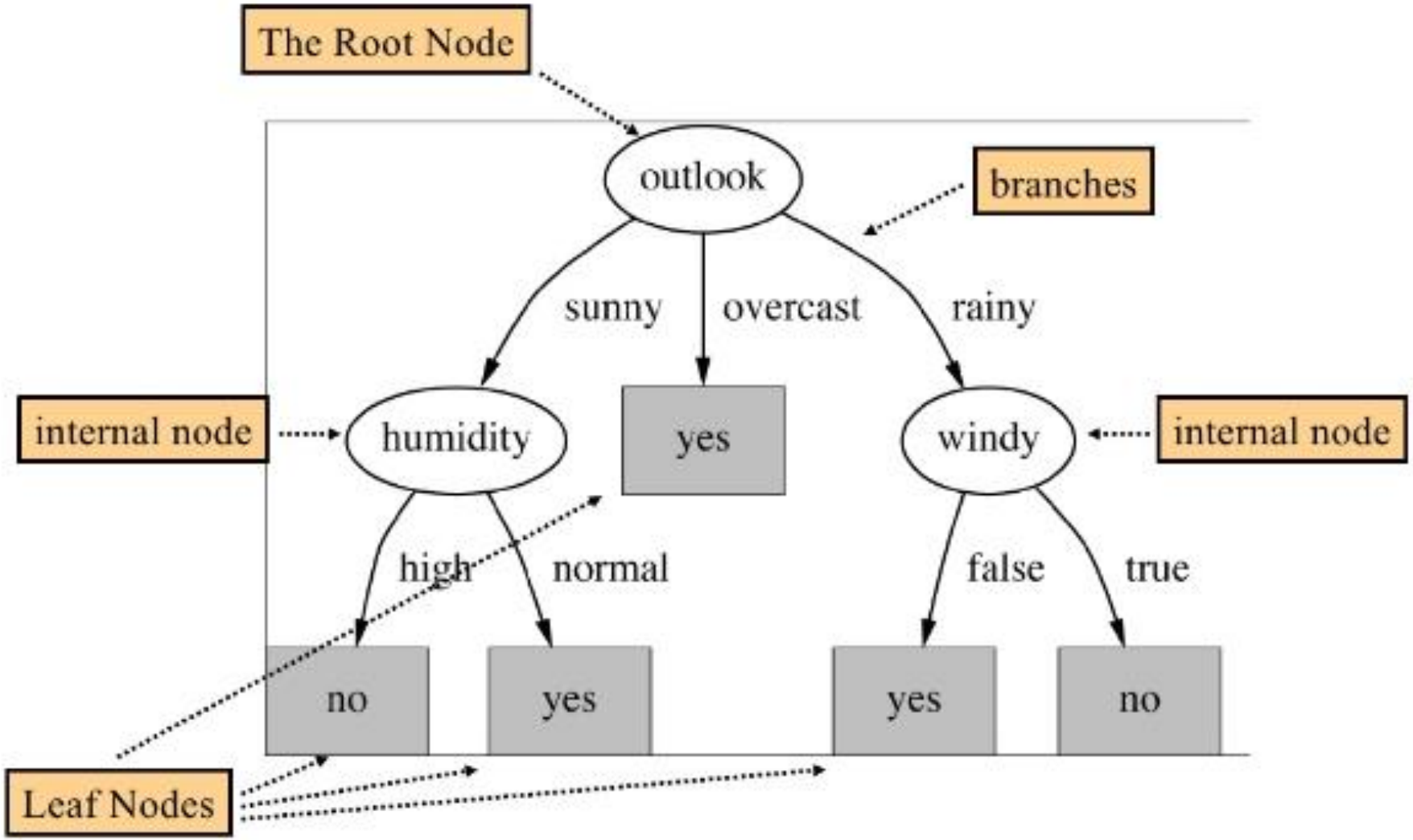
Decision Trees

Decision Tree Terminology

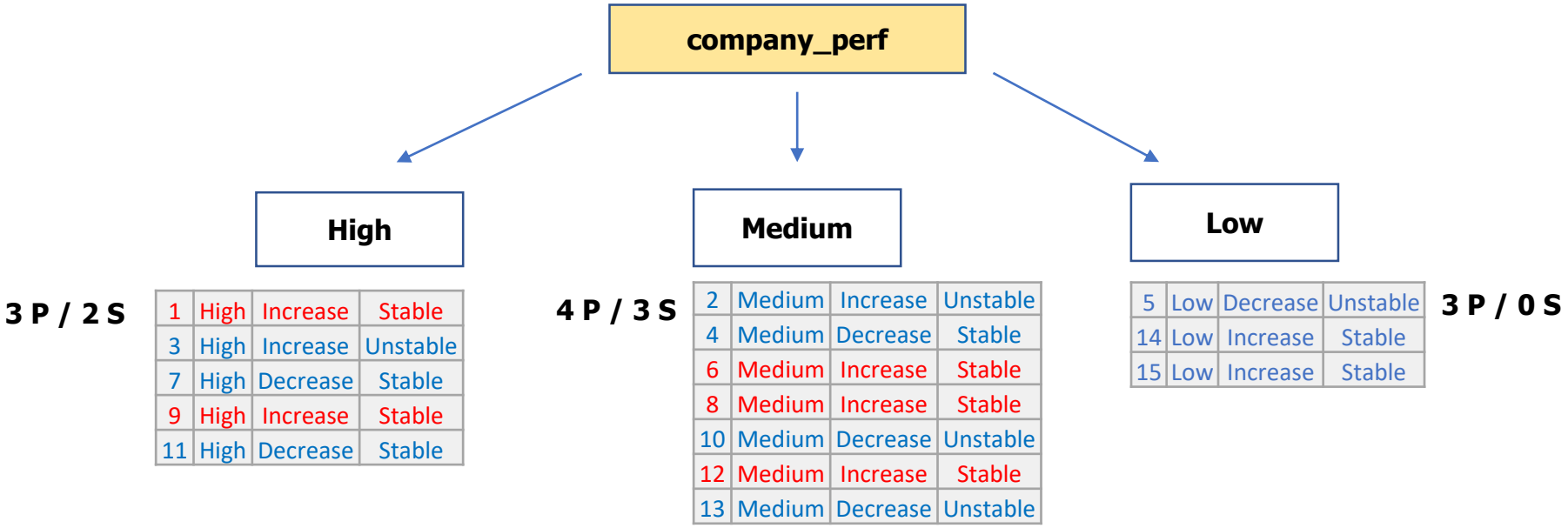


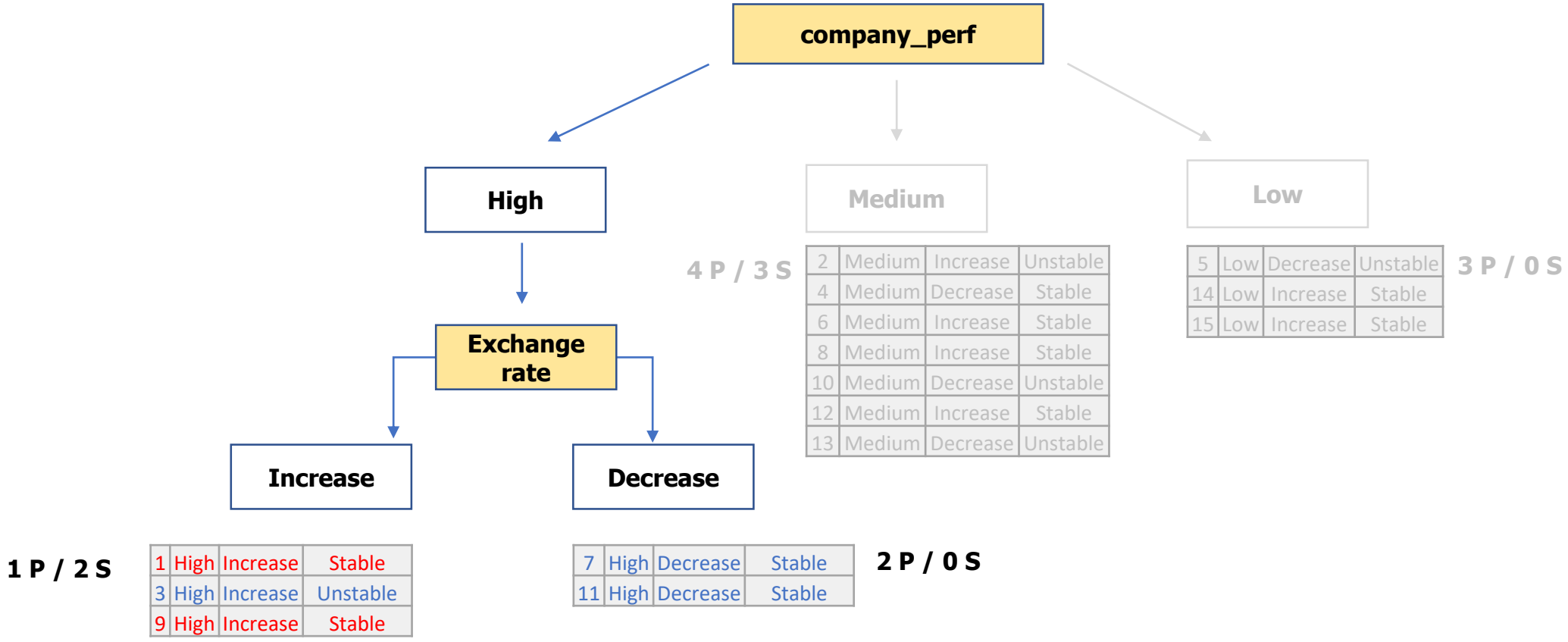
Predict whether someone will buy or sell stocks

Day	company_perf	exchange_rate	gold_price	Action
1	High	Increase	Stable	Sale
2	Medium	Increase	Unstable	Purchase
3	High	Increase	Unstable	Purchase
4	Medium	Decrease	Stable	Purchase
5	Low	Decrease	Unstable	Purchase
6	Medium	Increase	Stable	Sale
7	High	Decrease	Stable	Purchase
8	Medium	Increase	Stable	Sale
9	High	Increase	Stable	Sale
10	Medium	Decrease	Unstable	Purchase
11	High	Decrease	Stable	Purchase
12	Medium	Increase	Stable	Sale
13	Medium	Decrease	Unstable	Purchase
14	Low	Increase	Stable	Purchase
15	Low	Increase	Stable	Purchase
16	**Low**	**Decrease**	**Stable**	**????**

Company_perf High, Medium, Low
Exch_rate Increase, Decrease
Gold_price Stable, Unstable

Action 9 Purchase / 6 Sale


company_perf
High
Medium
Low
3 P / 2 S
1 High Increase Stable
3 High Increase Unstable
7 High Decrease Stable
9 High Increase Stable
11 High Decrease Stable
4 P / 3 S
2 Medium Increase Unstable
4 Medium Decrease Stable
6 Medium Increase Stable
8 Medium Increase Stable
10 Medium Decrease Unstable
12 Medium Increase Stable
13 Medium Decrease Unstable
5 Low Decrease Unstable
14 Low Increase Stable
15 Low Increase Stable
3 P / 0 S

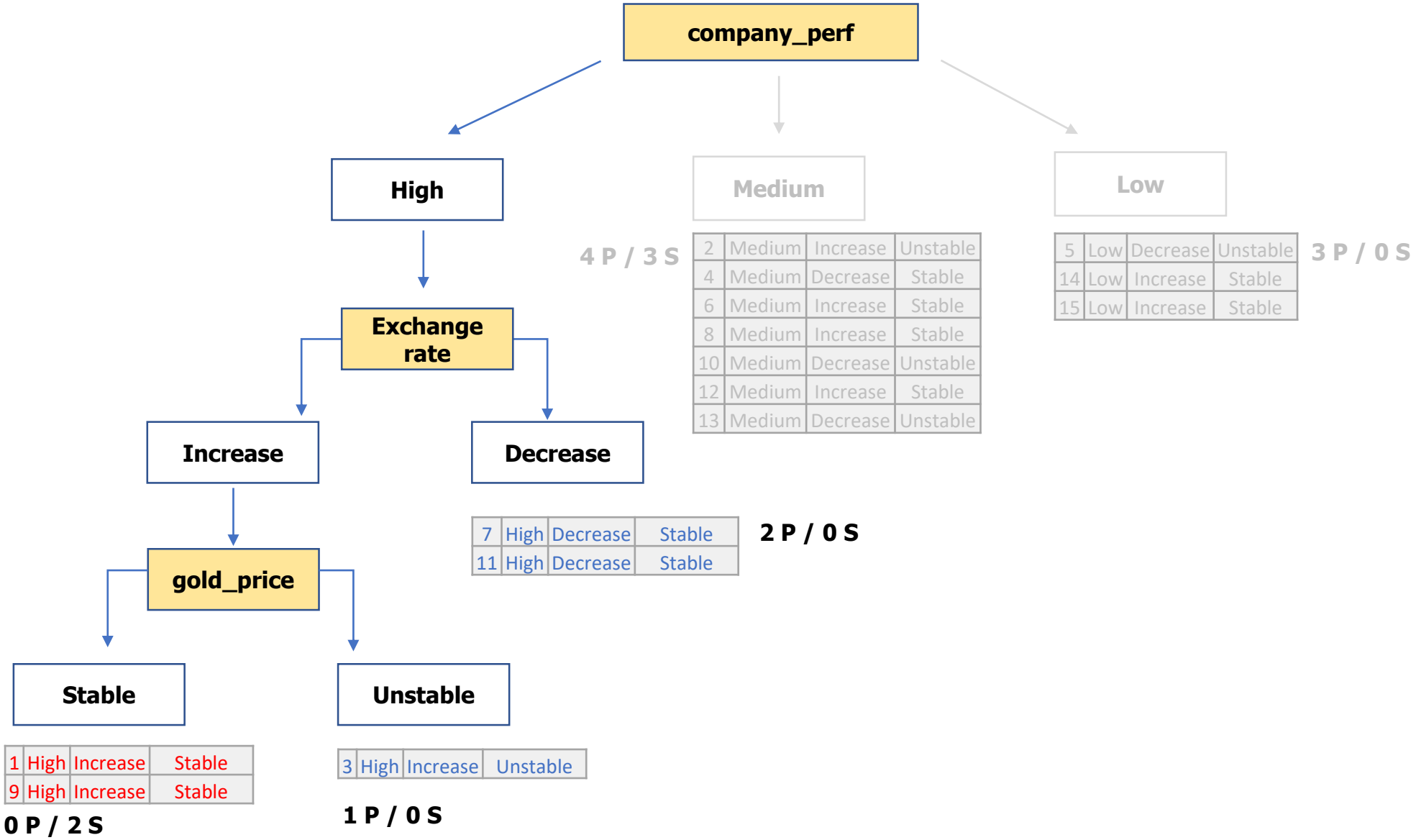


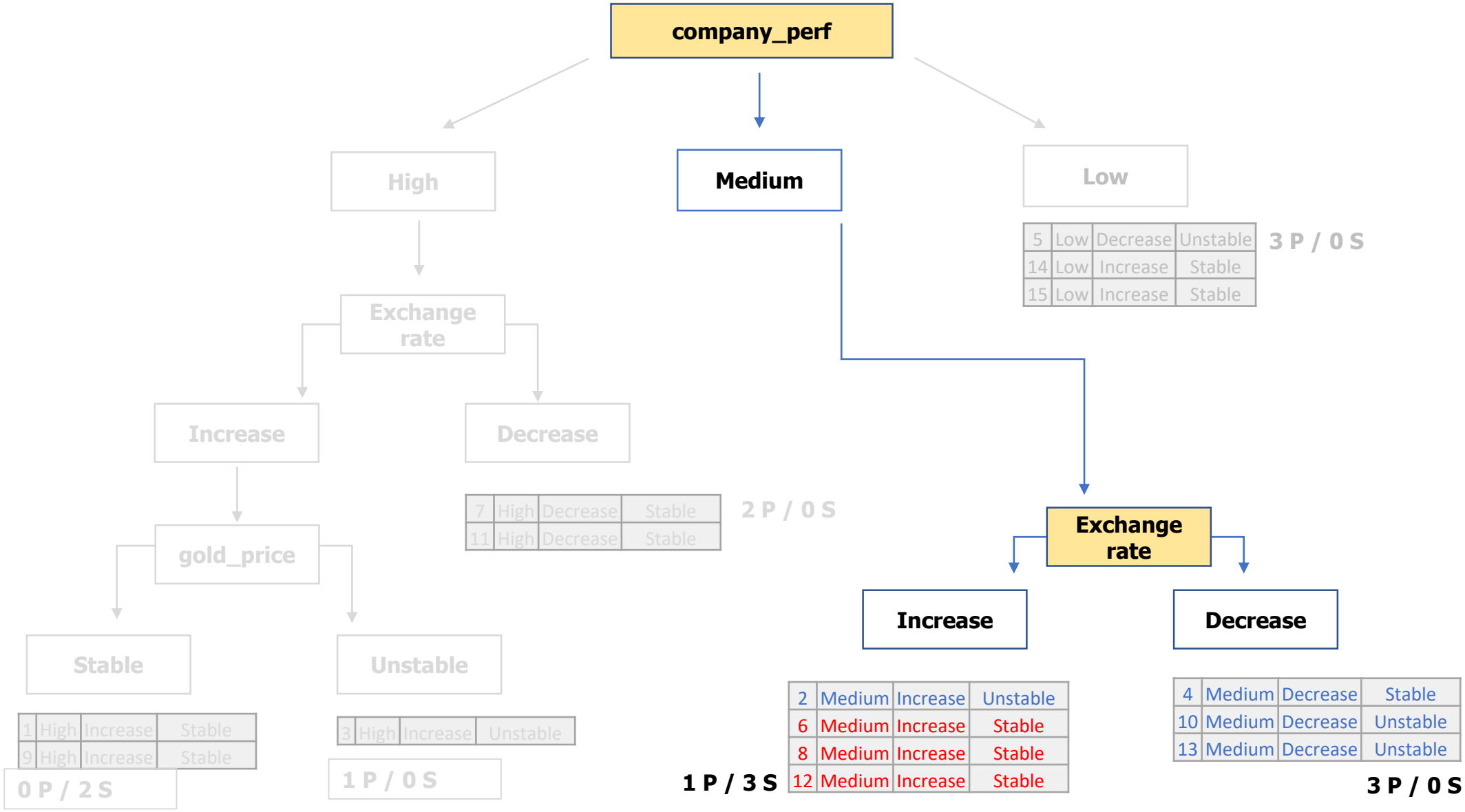
2	Medium	Increase	Unstable
4	Medium	Decrease	Stable
6	Medium	Increase	Stable
8	Medium	Increase	Stable
10	Medium	Decrease	Unstable
12	Medium	Increase	Stable
13	Medium	Decrease	Unstable

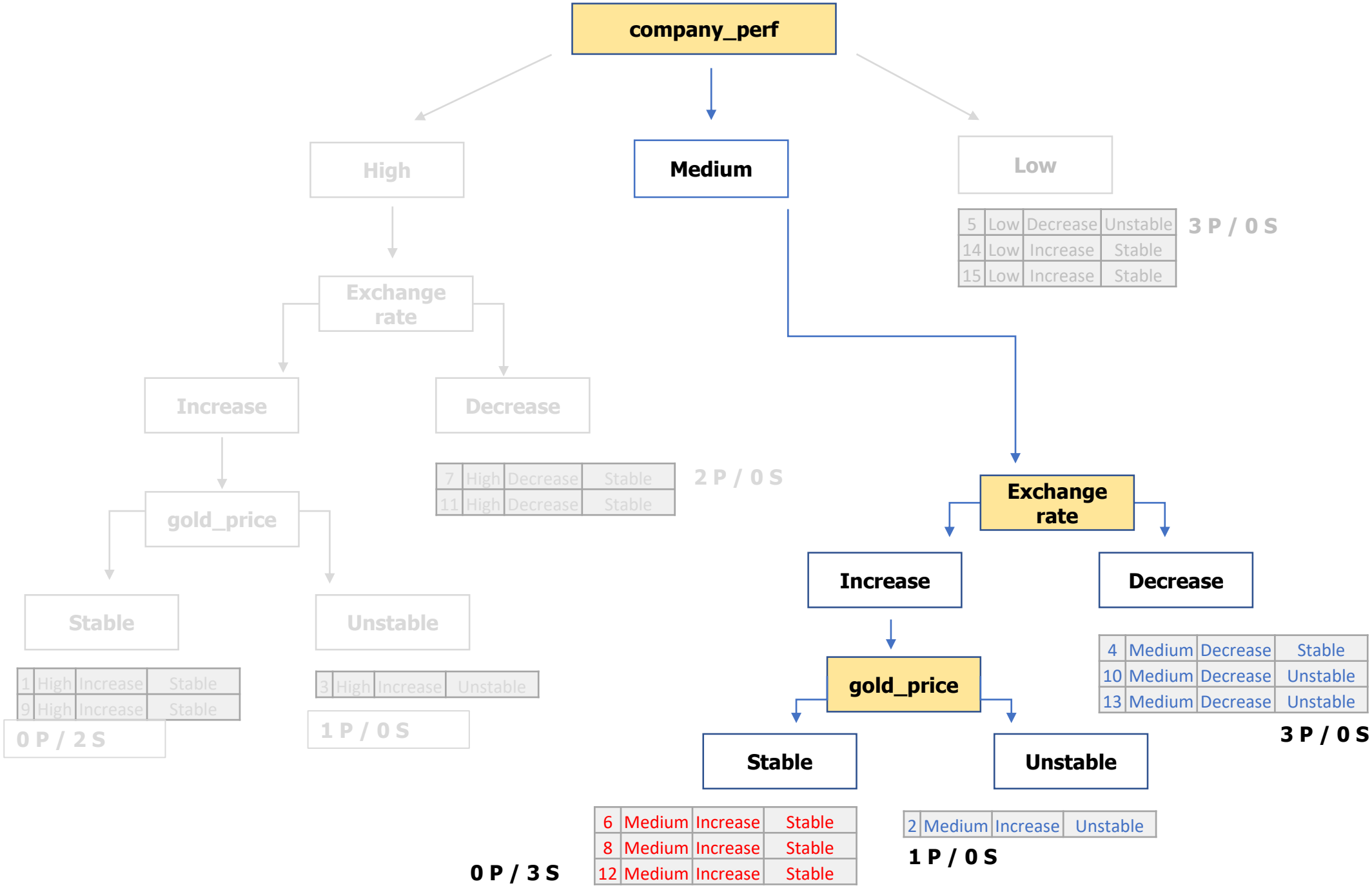
5	Low	Decrease	Unstable
14	Low	Increase	Stable
15	Low	Increase	Stable

1	High	Increase	Stable
3	High	Increase	Unstable
9	High	Increase	Stable

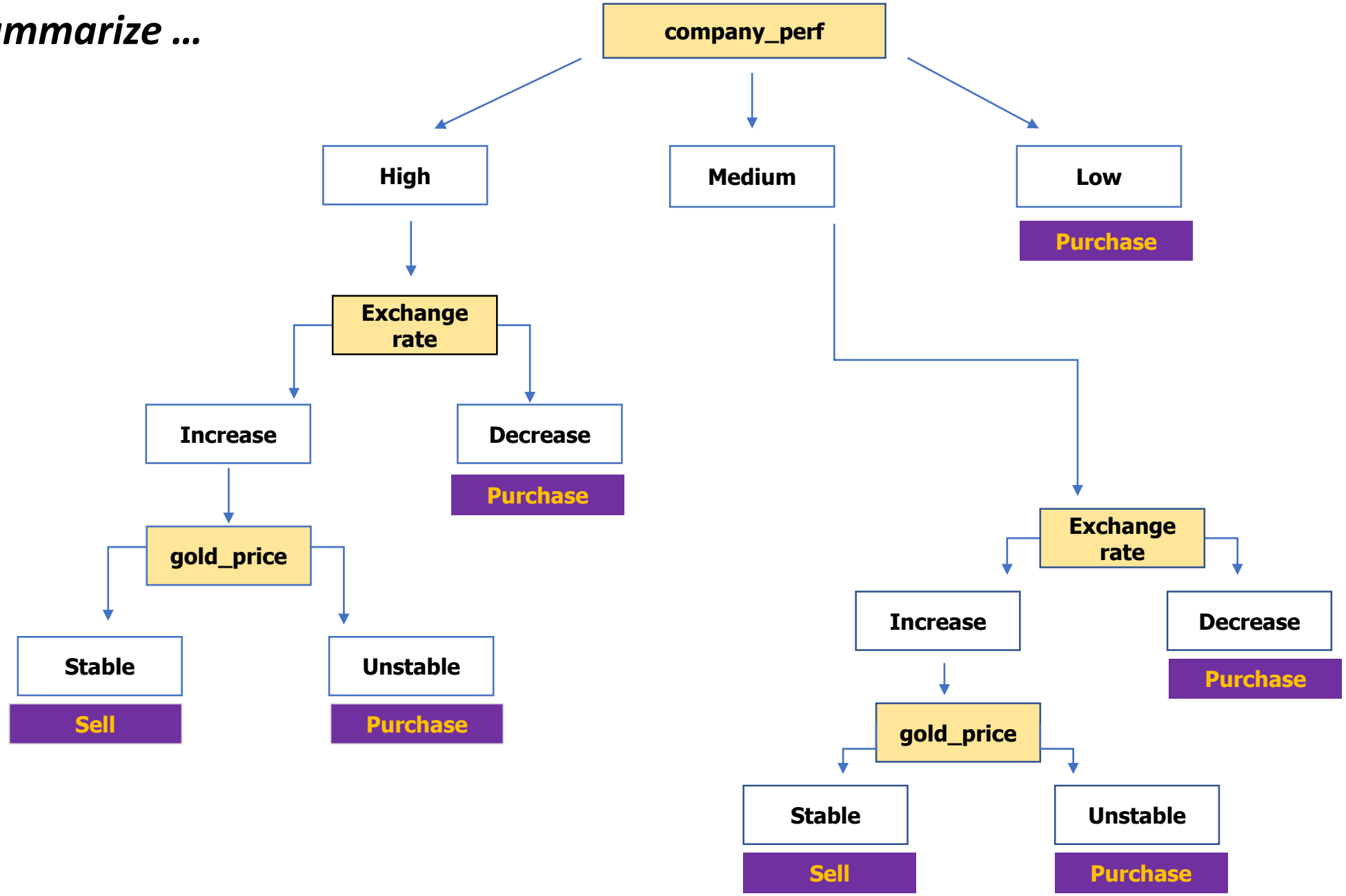
7	High	Decrease	Stable
11	High	Decrease	Stable

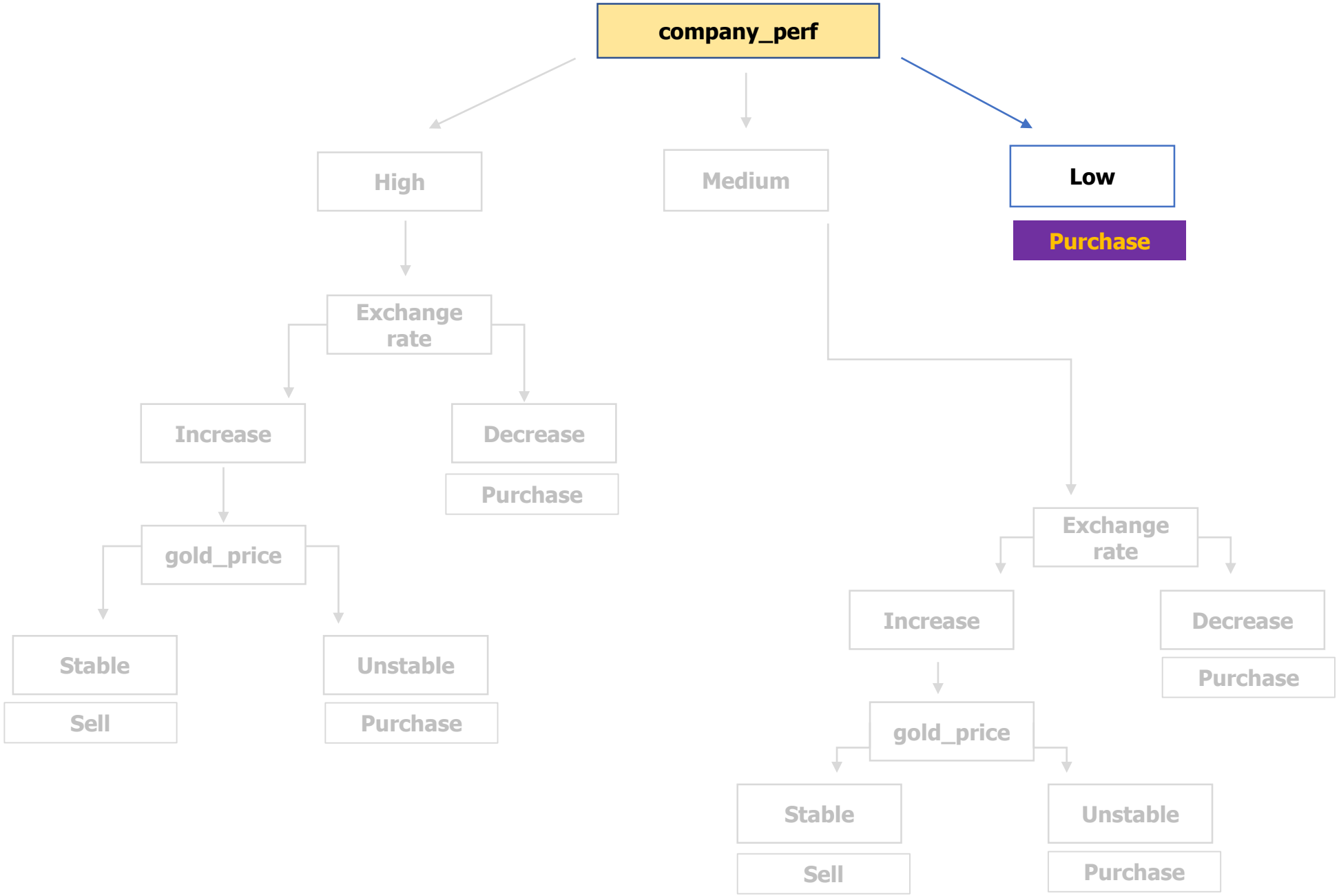

company_perf
High
Medium
Low
4 P / 3 S
2 Medium Increase Unstable
4 Medium Decrease Stable
6 Medium Increase Stable
8 Medium Increase Stable
10 Medium Decrease Unstable
12 Medium Increase Stable
13 Medium Decrease Unstable
5 Low Decrease Unstable
14 Low Increase Stable
15 Low Increase Stable
3 P / 0 S
Exchange rate
Increase
Decrease
7 High Decrease Stable
11 High Decrease Stable
2 P / 0 S
gold_price
Stable
Unstable
1 High Increase Stable
9 High Increase Stable
0 P / 2 S
3 High Increase Unstable
1 P / 0 S


company_perf
High
Medium
Low
5 Low Decrease Unstable
14 Low Increase Stable
15 Low Increase Stable
3 P / 0 S
Exchange rate
Increase
Decrease
7 High Decrease Stable
11 High Decrease Stable
2 P / 0 S
gold_price
Stable
Unstable
1 High Increase Stable
9 High Increase Stable
0 P / 2 S
3 High Increase Unstable
1 P / 0 S
Exchange rate
Increase
Decrease
2 Medium Increase Unstable
6 Medium Increase Stable
8 Medium Increase Stable
12 Medium Increase Stable
1 P / 3 S
4 Medium Decrease Stable
10 Medium Decrease Unstable
13 Medium Decrease Unstable
3 P / 0 S


company_perf
High
Medium
Low
5 Low Decrease Unstable
14 Low Increase Stable
15 Low Increase Stable
3 P / 0 S
Exchange rate
Increase
Decrease
7 High Decrease Stable
11 High Decrease Stable
2 P / 0 S
gold_price
Stable
Unstable
1 High Increase Stable
9 High Increase Stable
0 P / 2 S
3 High Increase Unstable
1 P / 0 S
Exchange rate
Increase
Decrease
4 Medium Decrease Stable
10 Medium Decrease Unstable
13 Medium Decrease Unstable
3 P / 0 S
gold_price
Stable
Unstable
6 Medium Increase Stable
8 Medium Increase Stable
12 Medium Increase Stable
0 P / 3 S
2 Medium Increase Unstable
1 P / 0 S

To summarize …





Day	company_perf	exchange_rate	gold_price	Action
16	Low	Decrease	Stable	????

Split criteria

Information Gain

- Significant variable to split is determined by **Information Gain**
- Measure that determines how well a given attribute splits the dataset
- This measure is used at every step to determine the next best attribute
- Information (I) is needed to classify an object

- **Gain (S, A) = E(S) – Σ [(S_a/S) * E(S_a)]**

where

E(S) = Entropy calculation

$\mathbf{S_a}$ = Dataset having attribute value **a**

S = Total count of dataset of attribute **A**

E(S_a) = Entropy of Attribute value **a**

- **Maximum**(Gain(A)) → Best Attribute

Entropy (I)

- Measures homogeneity of the sets
- Tells us how pure / impure a set is
- e.g. In a binary classification dataset, if **S** is the dataset having + and – classes, then Entropy (*I*nformation) is measured as:

$$\mathbf{E(S) = -p_{(+)}log_2\ p_{(+)} - p_{(-)}log_2 p_{(-)}}$$

where

$\mathbf{p_{(+)}}$ = % of positive class

$\mathbf{p_{(-)}}$ = % of negative class

- Interpretation of Entropy
  - ✓ 0 <= I <= 1
  - ✓ Number of bits that is needed to identify if an item in the given dataset is + or –
  - ✓ For a pure subset, number of bits = 0
  - ✓ For a tie, number of bits = 1

Refer to chart

Which attribute to split on?

10 P / 5 S

company_perf	Purchase	Sale
High	3	2
Low	3	0
Medium	4	3

company_perf	Purchase (+)	Sale (-)	PS_Total	$p_{(+)}$	$\log_2 p_{(+)}$	$-p_{(+)} \log_2 p_{(+)}$	$p_{(-)}$	$\log_2 p_{(-)}$	$p_{(-)} \log_2 p_{(-)}$	E
High	3	2	5	0.600	-0.737	0.442	0.400	-1.322	-0.529	0.971
Low	3	0	3	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Medium	4	3	7	0.571	-0.807	0.461	0.429	-1.222	-0.524	0.985
Set (Total)	10	5	15	0.667	-0.585	0.390	0.333	-1.585	-0.528	0.918

E(S)	S	$S_{(v=high)}$	$E(_{v=high})$	$(S_a/S) * E(S_a)$	$S_{(v=medium)}$	$E(_{v=medium})$	$(S_a/S) * E(S_a)$	Gain(S, company_perf)
0.918	15	5	0.971	0.324	7	0.985	0.460	0.783

10 P / 5 S

exchange_rate	Purchase	Sale
Decrease	6	0
Increase	4	5

exchange_rate	Purchase (+)	Sale (-)	PS_Total	$p_{(+)}$	$\log_2 p_{(+)}$	$-p_{(+)} \log_2 p_{(+)}$	$p_{(-)}$	$\log_2 p_{(-)}$	$p_{(-)} \log_2 p_{(-)}$	E
Decrease	6	0	6	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Increase	4	5	9	0.444	-1.170	0.520	0.556	-0.848	-0.471	0.991
Set (Total)	10	5	15	0.667	-0.585	0.390	0.333	-1.585	-0.528	0.918

E(S)	S	$S_{(v=decrease)}$	$E(_{v=decrease})$	$(S_a/S) * E(S_a)$	$S_{(v=increase)}$	$E(_{v=increase})$	$(S_a/S) * E(S_a)$	Gain(S, exchange_rate)
0.918	15	6	0.000	0.000	9	0.991	0.595	0.595

10 P / 5 S

gold_price	Purchase	Sale
Stable	5	5
Unstable	5	0

gold_price	Purchase (+)	Sale (-)	PS_Total	$p_{(+)}$	$\log_2 p_{(+)}$	$-p_{(+)} \log_2 p_{(+)}$	$p_{(-)}$	$\log_2 p_{(-)}$	$p_{(-)} \log_2 p_{(-)}$	E
stable	5	5	10	0.500	-1.000	0.500	0.500	-1.000	-0.500	1.000
unstable	5	0	5	1.000	0.000	0.000	0.000	0.000	0.000	0.000
Set (Total)	10	5	15	0.667	-0.585	0.390	0.333	-1.585	-0.528	0.918

E(S)	S	$S_{(v=stable)}$	$E(_{v=stable})$	$(S_a/S) * E(S_a)$	$S_{(v=unstable)}$	$E(_{v=medium})$	$(S_a/S) * E(S_a)$	Gain(S, company_perf)
0.918	15	10	1.000	0.667	5	0.000	0.000	0.667

Pure Impure

Click here to view the detailed calculations