TON DUC THANG UNIVERSITY
Faculty of Information Technology

DECISION TREE AND ASSOCIATION RULE LEARNING

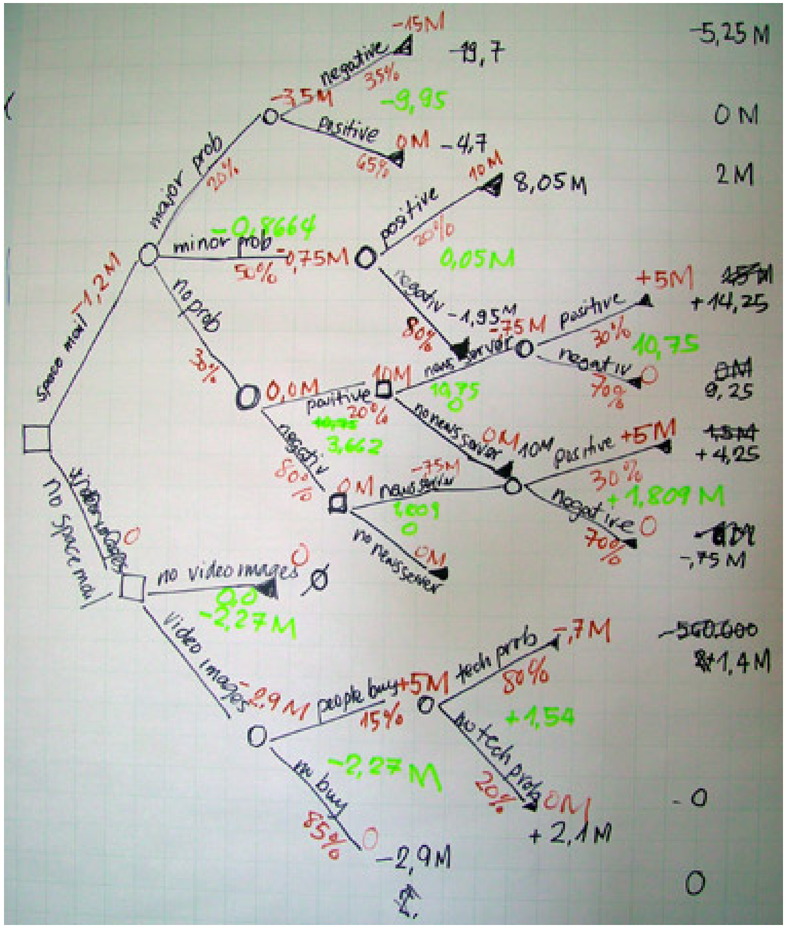
Mid-term project of **Data Mining and Knowledge Discovery (505043)**

Huynh Van Duy (51703006) – **Tran Quoc Linh** (51703124)
Adviser **Tran Thanh Phuoc (Ph. D.)**

Ho Chi Minh, October 5-th, 2019

DECISION TREE

- A **decision tree** is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that **only contains conditional control statements.**

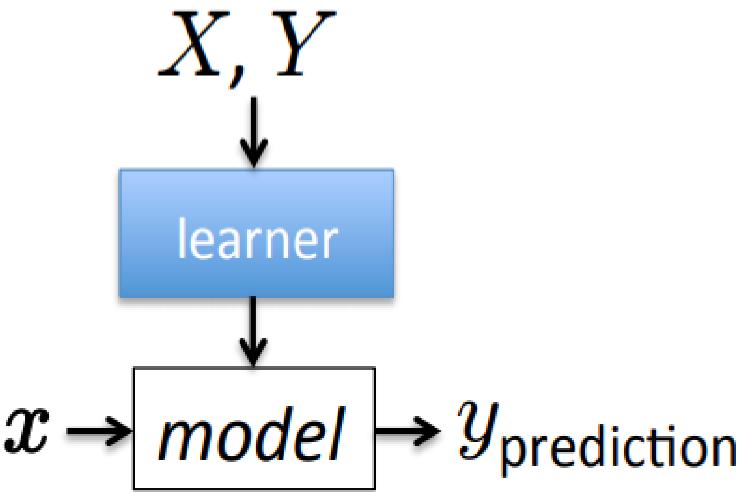


DECISION TREE

- **Given:** labeled training data $X, Y = \{\langle \boldsymbol{x}_i, y_i \rangle\}_{i=1}^{n}$
  - Assumes each $\boldsymbol{x}_i \sim \mathcal{D}(\mathcal{X})$ with $y_i = f_{target}(\boldsymbol{x}_i)$
- **Train model:**
  - **model** := classifier.train(X, Y)
- **Test:** new unlabeled instance

$$\boldsymbol{x} \sim \mathcal{D}(\mathcal{X})$$

y_prediction := model.predict(x)



DECISION TREE

- **Information gain** tells us how important a given attribute of the features is.
- Use this information to decide the ordering of attributes in the nodes of a decision tree:

$$Gain(S,A) = I_S(A,Y) = H_S(Y) - H_S(Y|A)$$

- Entropy of a random variable:

$$H(X) = -\sum_{i=1}^{n} P(X=i)\log_2 P(X=i)$$

- Conditional entropy of X given Y:

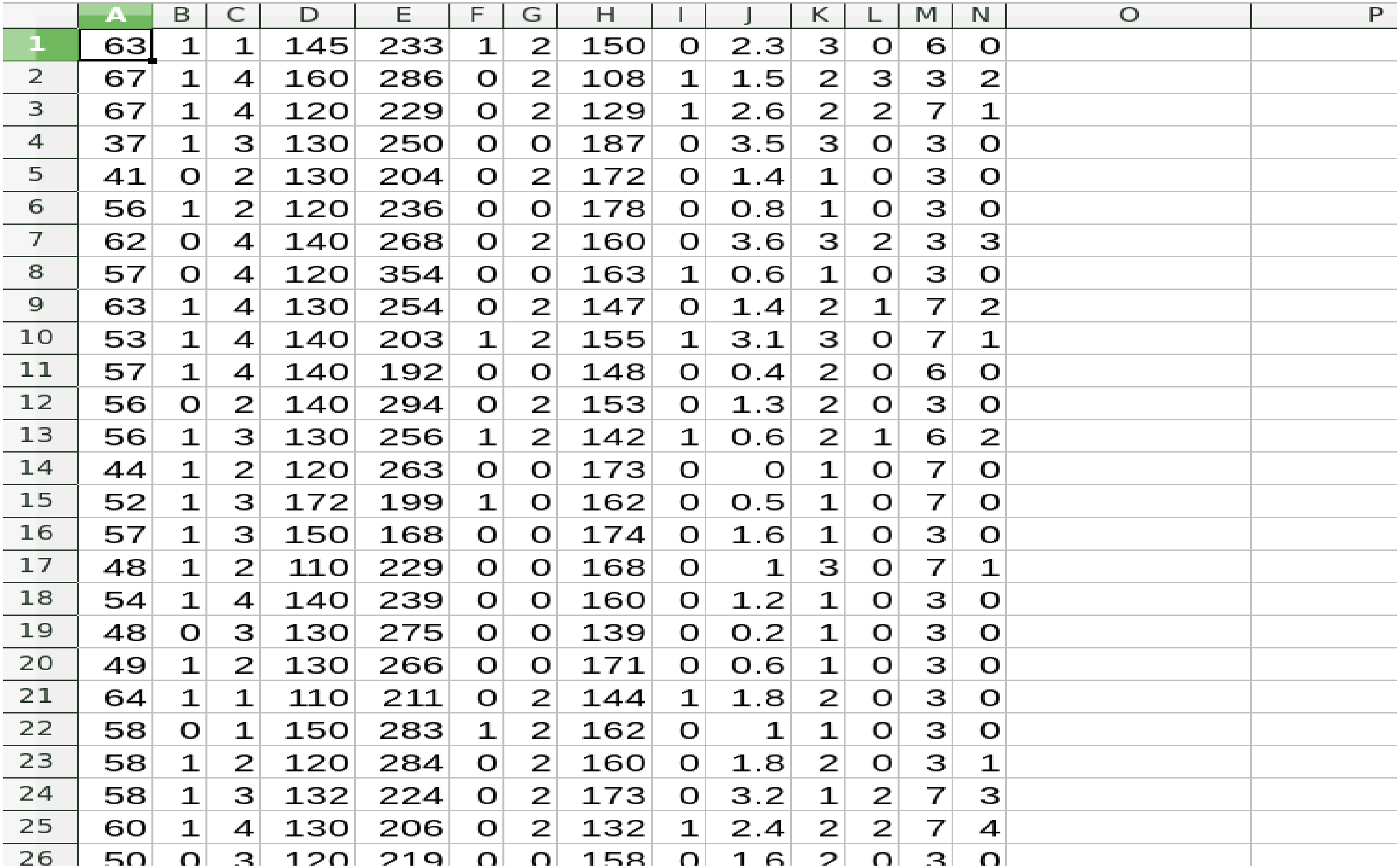
$$H(X|Y) = \sum_{v \in values(Y)} P(Y=v)H(X|Y=v)$$

DECISION TREE: EXPERIENCE

- **Heart Disease UCI**: This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

- 14 attributes: age, sex, chest pain type, resting blood pressure, serum cholesterol (mg/dl), fasting blood sugar (> 120 mg/dl), resting electrocardiographic results (0, 1, 2), maximum heart rate achieved, exercise included angina, oldpeak, the slope of the peak exercise ST segment, number of major vessels (0-3), thal (3 = normal, 6 = fixed defect, 7 = reversable defect) and target which refer to the type of heart disease (0-4).

DECISION TREE: EXPERIENCE

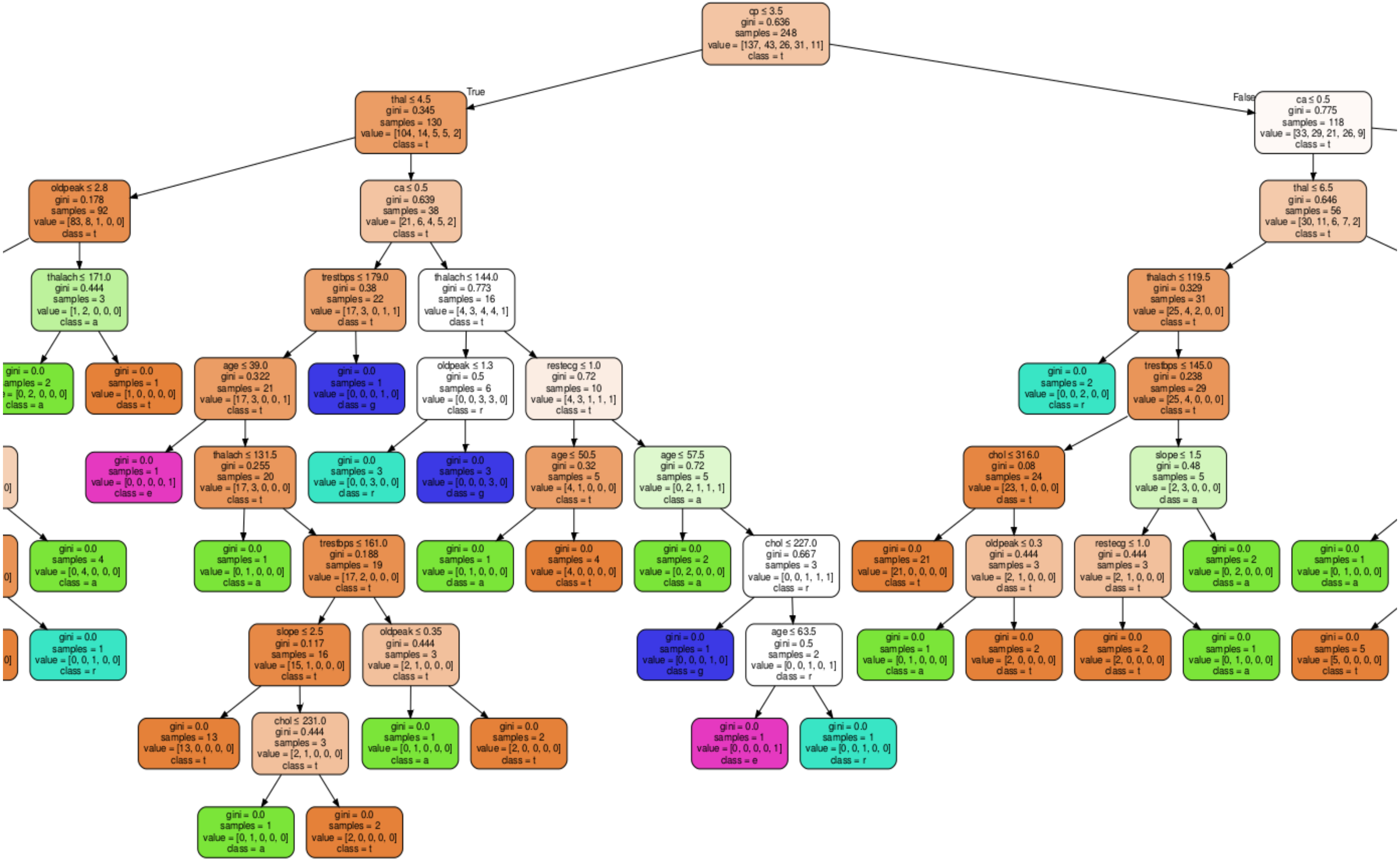


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0		
2	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2		
3	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1		
4	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0		
5	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0		
6	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0		
7	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3		
8	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0		
9	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2		
10	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1		
11	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0		
12	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0		
13	56	1	3	130	256	1	2	142	1	0.6	2	1	6	2		
14	44	1	2	120	263	0	0	173	0	0	1	0	7	0		
15	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0		
16	57	1	3	150	168	0	0	174	0	1.6	1	0	3	0		
17	48	1	2	110	229	0	0	168	0	1	3	0	7	1		
18	54	1	4	140	239	0	0	160	0	1.2	1	0	3	0		
19	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0		
20	49	1	2	130	266	0	0	171	0	0.6	1	0	3	0		
21	64	1	1	110	211	0	2	144	1	1.8	2	0	3	0		
22	58	0	1	150	283	1	2	162	0	1	1	0	3	0		
23	58	1	2	120	284	0	2	160	0	1.8	2	0	3	1		
24	58	1	3	132	224	0	2	173	0	3.2	1	2	7	3		
25	60	1	4	130	206	0	2	132	1	2.4	2	2	7	4		
26	50	0	3	120	219	0	0	158	0	1.6	2	0	3	0		

DECISION TREE: EXPERIENCE

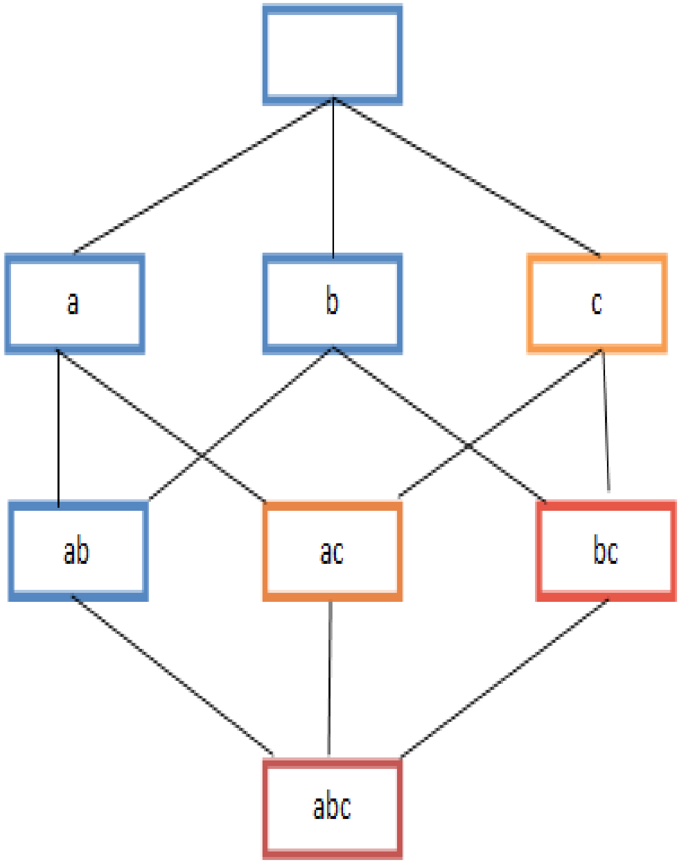
- This database contains 297 samples. We split this database into 2 subset: train (248) and test (49).
- Library: **sklearn.tree.DecisionTreeClassifier.**
- Result:
  - Train accuracy: 53%
  - Test accuracy: 49%

DECISION TREE: EXPERIENCE


cp ≤ 3.5 gini = 0.636 samples = 248 value = [137, 43, 26, 31, 11] class = t
True
False
thal ≤ 4.5 gini = 0.345 samples = 130 value = [104, 14, 5, 5, 2] class = t
ca ≤ 0.5 gini = 0.775 samples = 118 value = [33, 29, 21, 26, 9] class = t
oldpeak ≤ 2.8 gini = 0.178 samples = 92 value = [83, 8, 1, 0, 0] class = t
ca ≤ 0.5 gini = 0.639 samples = 38 value = [21, 6, 4, 5, 2] class = t
thal ≤ 6.5 gini = 0.646 samples = 56 value = [30, 11, 6, 7, 2] class = t
thalach ≤ 171.0 gini = 0.444 samples = 3 value = [1, 2, 0, 0, 0] class = a
trestbps ≤ 179.0 gini = 0.38 samples = 22 value = [17, 3, 0, 1, 1] class = t
thalach ≤ 144.0 gini = 0.773 samples = 16 value = [4, 3, 4, 4, 1] class = t
thalach ≤ 119.5 gini = 0.329 samples = 31 value = [25, 4, 2, 0, 0] class = t
gini = 0.0 amples = 2 = [0, 2, 0, 0, 0] class = a
gini = 0.0 samples = 1 value = [1, 0, 0, 0, 0] class = t
age ≤ 39.0 gini = 0.322 samples = 21 value = [17, 3, 0, 0, 1] class = t
gini = 0.0 samples = 1 value = [0, 0, 0, 1, 0] class = g
oldpeak ≤ 1.3 gini = 0.5 samples = 6 value = [0, 0, 3, 3, 0] class = r
restecg ≤ 1.0 gini = 0.72 samples = 10 value = [4, 3, 1, 1, 1] class = t
gini = 0.0 samples = 2 value = [0, 0, 2, 0, 0] class = r
trestbps ≤ 145.0 gini = 0.238 samples = 29 value = [25, 4, 0, 0, 0] class = t
0]
gini = 0.0 samples = 1 value = [0, 0, 0, 0, 1] class = e
thalach ≤ 131.5 gini = 0.255 samples = 20 value = [17, 3, 0, 0, 0] class = t
gini = 0.0 samples = 3 value = [0, 0, 3, 0, 0] class = r
gini = 0.0 samples = 3 value = [0, 0, 0, 3, 0] class = g
age ≤ 50.5 gini = 0.32 samples = 5 value = [4, 1, 0, 0, 0] class = t
age ≤ 57.5 gini = 0.72 samples = 5 value = [0, 2, 1, 1, 1] class = a
chol ≤ 316.0 gini = 0.08 samples = 24 value = [23, 1, 0, 0, 0] class = t
slope ≤ 1.5 gini = 0.48 samples = 5 value = [2, 3, 0, 0, 0] class = a
0]
gini = 0.0 samples = 4 value = [0, 4, 0, 0, 0] class = a
gini = 0.0 samples = 1 value = [0, 1, 0, 0, 0] class = a
trestbps ≤ 161.0 gini = 0.188 samples = 19 value = [17, 2, 0, 0, 0] class = t
gini = 0.0 samples = 1 value = [0, 1, 0, 0, 0] class = a
gini = 0.0 samples = 4 value = [4, 0, 0, 0, 0] class = t
gini = 0.0 samples = 2 value = [0, 2, 0, 0, 0] class = a
chol ≤ 227.0 gini = 0.667 samples = 3 value = [0, 0, 1, 1, 1] class = r
gini = 0.0 samples = 21 value = [21, 0, 0, 0, 0] class = t
oldpeak ≤ 0.3 gini = 0.444 samples = 3 value = [2, 1, 0, 0, 0] class = t
restecg ≤ 1.0 gini = 0.444 samples = 3 value = [2, 1, 0, 0, 0] class = t
gini = 0.0 samples = 2 value = [0, 2, 0, 0, 0] class = a
gini = 0.0 samples = 1 value = [0, 1, 0, 0, 0] class = a
0]
gini = 0.0 samples = 1 value = [0, 0, 1, 0, 0] class = r
slope ≤ 2.5 gini = 0.117 samples = 16 value = [15, 1, 0, 0, 0] class = t
oldpeak ≤ 0.35 gini = 0.444 samples = 3 value = [2, 1, 0, 0, 0] class = t
gini = 0.0 samples = 1 value = [0, 0, 0, 1, 0] class = g
age ≤ 63.5 gini = 0.5 samples = 2 value = [0, 0, 1, 0, 1] class = r
gini = 0.0 samples = 1 value = [0, 1, 0, 0, 0] class = a
gini = 0.0 samples = 2 value = [2, 0, 0, 0, 0] class = t
gini = 0.0 samples = 2 value = [2, 0, 0, 0, 0] class = t
gini = 0.0 samples = 1 value = [0, 1, 0, 0, 0] class = a
gini = 0.0 samples = 5 value = [5, 0, 0, 0, 0] class = t
gini = 0.0 samples = 13 value = [13, 0, 0, 0, 0] class = t
chol ≤ 231.0 gini = 0.444 samples = 3 value = [2, 1, 0, 0, 0] class = t
gini = 0.0 samples = 1 value = [0, 1, 0, 0, 0] class = a
gini = 0.0 samples = 2 value = [2, 0, 0, 0, 0] class = t
gini = 0.0 samples = 1 value = [0, 0, 0, 0, 1] class = e
gini = 0.0 samples = 1 value = [0, 0, 1, 0, 0] class = r
gini = 0.0 samples = 1 value = [0, 1, 0, 0, 0] class = a
gini = 0.0 samples = 2 value = [2, 0, 0, 0, 0] class = t

ASSOCIATION RULE LEARNING

- **Association rule learning** is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.



ASSOCIATION RULE LEARNING

- **Support** is an indication of how frequently the itemset appears in the dataset. The support of **X** with respect to **T** is defined as the proportion of transactions **t** in the dataset which contains the itemset **X**:

$$\mathrm{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

- **Condidence** is an indication of how often the rule has been found to be true.The confidence value of a rule, **X → Y** , with respect to a set of transactions **T**, is the proportion of the transactions that contains **X** which also contains **Y**.

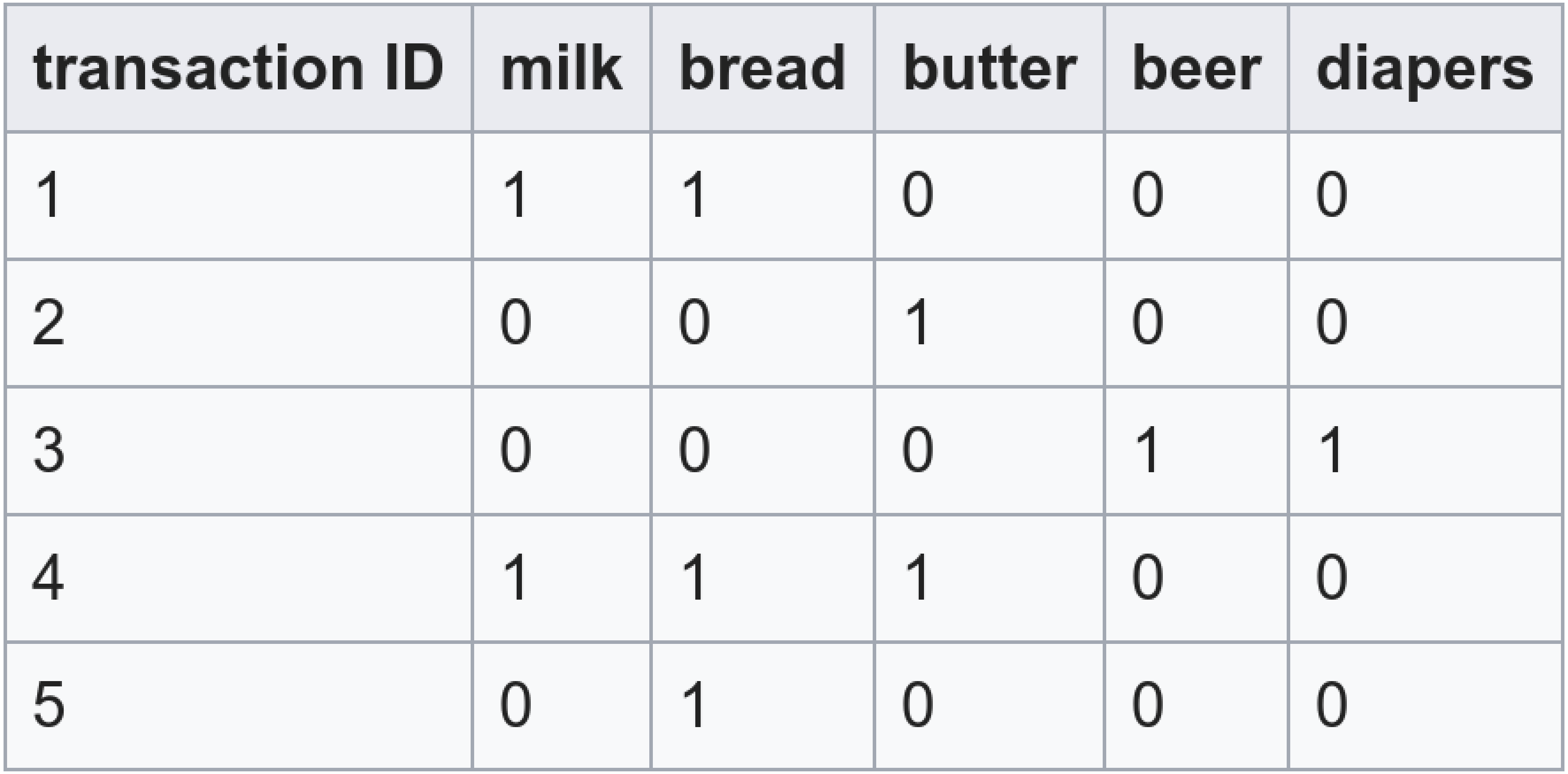
$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

ASSOCIATION RULE LEARNING

- Many algorithms for generating association rules have been proposed:
  - **Apriori alogrithm** uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support.
  - **Eclat algorithm** is a depth-first search algorithm based on set intersection. It is suitable for both sequential as well as parallel execution with locality-enhancing properties.
  - **FP-growth algorithm** is an improvement of apriori algorithm. FP growth algorithm used for finding frequent itemset in a transaction database without candidate generation.
  - ...

ASSOCIATION RULE LEARNING: EXPERIENCE

- Given list of transactions with unique items. Find all association rules, which satisfy with the given min support and min confidence.



transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

THANKS FOR YOUR ATTENTION!