Machine Learning with Python

Classification and Regression Trees

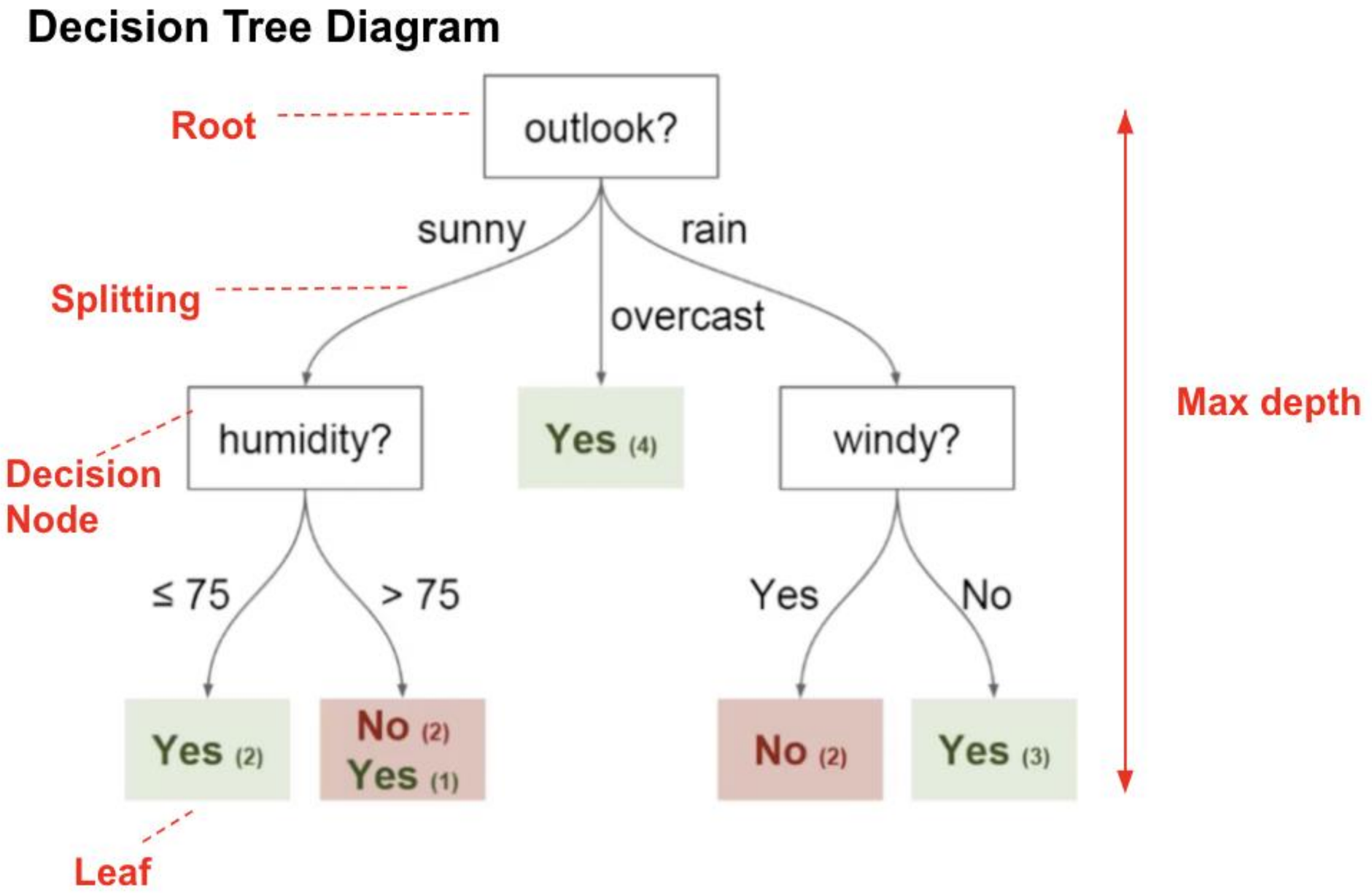
Arghya Ray

Decision Tree

- A decision tree is a popular classification method that results in a flow-chart like tree structure where each node denotes a test on an attribute value and each branch represents an outcome of the test. The tree leaves represent the classes.
- Decision tree is a model that is both predictive and descriptive.

- **Advantages**:
  - Decision tree approach is widely used since it is efficient and can deal with both continuous and categorical variables.
  - The decision tree approach is able to deal with missing values in the training data and can tolerate some errors in data.
  - The decision tree approach is perhaps the best if each attribute takes only a small number of possible values.
- **Disadvantages**:
  - Decision trees are less appropriate for tasks where the task is to predict values of a continuous variable like share price or interest rate.
  - Decision trees can lead to a large number of errors if the number of training examples per class is small.
  - The complexity of a decision tree increases as the number of attributes increases.

- ***Measuring the quality of a decision tree*** is an interesting problem altogether. ***Classification accuracy*** determined using test data is obviously a good measure but other measures like, ***average cost*** and ***worst case cost*** of classifying an object may be used.

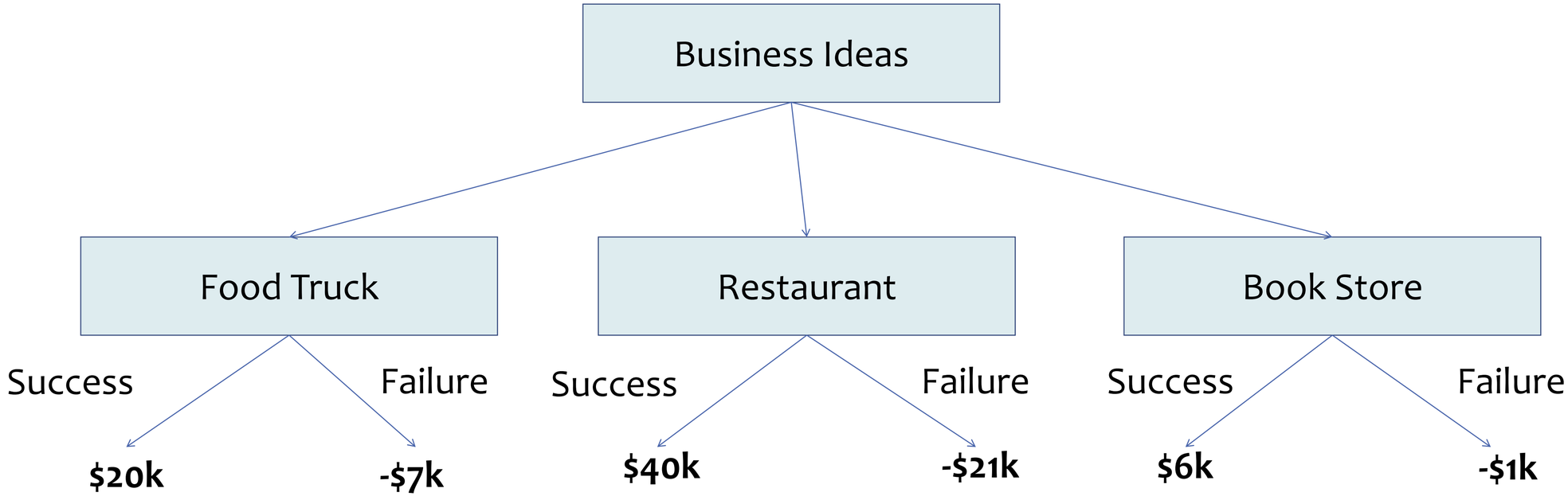


Picture taken from Velocity Business Solutions. Link: https://www.vebuso.com/2020/01/decision-tree-intuition-from-concept-to-application/

1. A decision tree is an approach to analysis that can help you make decisions.

Suppose for example you need to decide whether to invest a certain amount of money in one of the three business projects: a food-truck business, a restaurant, or a bookstore based on the data given below.

	Business Success Percentage		Business Value Changes	
Business	**Success Rate**	**Failure Rate**	**Gain (USD)**	**Loss (USD)**
Food Truck	60%	40%	20000	-7000
Restaurant	52%	48%	40000	-21000
Bookstore	50%	50%	6000	-1000



- In these cases, ***the expected value*** calculated based on all possible outcomes helps in figuring out the business decision making.

- Expected Value for the food truck business = (60% of USD 20000)+ (40% of USD (-7000)) = USD 9200.
- Expected Value of restaurant business = (52% of USD 40000) + (48% of USD (-21000)) = USD 10720.
- Expected Value of bookstore business = (50% of USD 6000) + (50% of USD (-1000)) = USD 2500

- Here the expected value reflects the average gain from investing in the business. Based on the above hypothetical figures, the results reflect that if you attempt to invest in a businesses say Food Truck business several times (under the same circumstances each time), your average profit will be USD 9200 per business.

2. Decision trees can also be used to visualize classification rules.

Classification and Regression Trees

Goal: Classify or predict an outcome based on a set of predictors.

The output is a set of **rules**

Example:

- Goal: classify a record as "will accept credit card offer" or "will not accept"
- Rule might be "IF (Income > 92.5) AND (Education < 1.5) AND (Family <= 2.5) THEN Class = 0 (non-acceptor)
- **Recursive partitioning:** Repeatedly split the records into two parts so as to achieve maximum homogeneity within the new parts

Recursive partitioning steps:

- Pick one of the predictor variables, xi
- Pick a value of xi, say si, that divides the training data into two (not necessarily equal) portions
- Measure how "pure" or homogeneous each of the resulting portions are
- "Pure" = containing records of mostly one class
- Algorithm tries with different variables (x) and different values of x_i, , i.e., s_i to maximize purity in a split
- After you get a "maximum purity" split, repeat the process for a second split, and so on

Forming a tree from the given example

RID	Age	Income	Student	Credit rating	Class (buys computer)
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31-40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31-40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Excellent	Yes
11	<=30	Medium	Yes	Excellent	Yes
12	30-40	Medium	No	Excellent	Yes
13	30-40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

<=30

income	student	Credit rating	Class
High	No	fair	No
High	No	excellent	No
Medium	No	Fair	No
Low	Yes	Fair	Yes
Medium	Yes	Excellent	yes

31-40

income	student	Credit rating	Class
High	No	Fair	Yes
Low	Yes	Excellent	Yes
Medium	No	Excellent	Yes
High	Yes	Fair	Yes

>40

income	student	Credit rating	Class
Medium	No	fair	Yes
Low	Yes	Fair	Yes
Low	Yes	Excellent	No
Medium	Yes	Fair	Yes
Medium	No	Excellent	No

Measuring Impurity

- Gini Index (measure of impurity)
  - Gini Index for rectangle *A* containing *m* cases

$$I(A) = 1 - \sum_{k=1}^{m} p_k^2$$

p = proportion of cases in rectangle *A* that belong to class *k*

  - I(A) = 0 when all cases belong to same class (most pure)

- Entropy (measure of impurity)

$$entropy\,(A) = -\sum_{k=1}^{m} p_k \log_2(p_k)$$

p = proportion of cases (out of *m*) in rectangle *A* that belong to class *k*

Entropy ranges between 0 (most pure) and $\log_2(m)$ (equal representation of classes)

Using the principle of ‘Information entropy’ build a ‘decision tree’ using the training data given below. Divide the ‘credit rating’ attribute into ranges as follows: (0, 1.6], (1.6,1.7], (1.7,1.8], (1.8,1.9], (1.9,2.0], (2.0,5.0]

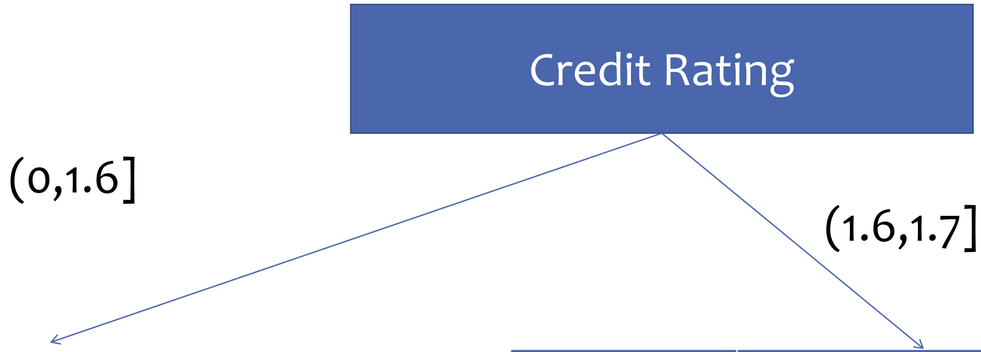
Sr. No.	Profession	Credit rating	Class
1	Business	1.6	Buys only laptop
2	Service	2.0	Buys laptop with CD Writer
3	Business	1.9	Buys laptop with printer
4	Business	1.88	Buys laptop with printer
5	Business	1.70	Buys only laptop
6	Service	1.85	Buys laptop with printer
7	Business	1.60	Buys only laptop
8	Service	1.70	Buys only laptop
9	Service	2.20	Buys laptop with CD writer
10	Service	2.10	Buys laptop with CD writer
11	Business	1.80	Buys laptop with printer
12	Service	1.95	Buys laptop with printer
13	Business	1.90	Buys laptop with printer
14	Business	1.80	Buys laptop with printer
15	Business	1.75	Buys laptop with printer

Business

Service

Sr. No.	Credit rating	Class
1	1.6	Buys only laptop
2	1.9	Buys laptop with printer
3	1.88	Buys laptop with printer
4	1.70	Buys only laptop
5	1.60	Buys only laptop
6	1.80	Buys laptop with printer
7	1.90	Buys laptop with printer
8	1.80	Buys laptop with printer
9	1.75	Buys laptop with printer

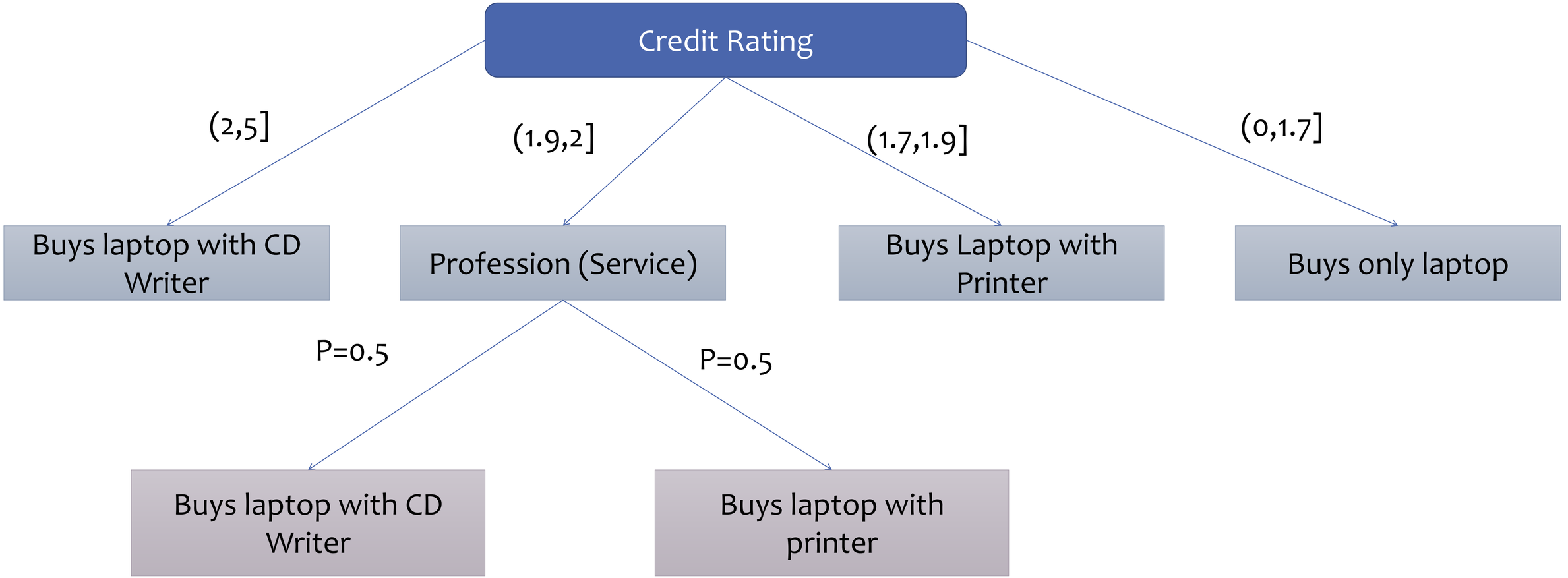
Sr. No.	Credit rating	Class
1	2.0	Buys laptop with CD Writer
2	1.85	Buys laptop with printer
3	1.70	Buys only laptop
4	2.20	Buys laptop with CD writer
5	2.10	Buys laptop with CD writer
6	1.95	Buys laptop with printer



Sr. No.	Profession	Class
1	Business	Buys only Laptop
2	Business	Buys only Laptop

Sr. No.	Profession	Class
1	Business	Buys only Laptop
2	Service	Buys only Laptop

- Initially there are 3 classes: Buys only laptop, buys laptop with CD writer, buys laptop with printer
- Initial Overall Entropy (E_0)= $-\sum_{i=1}^{3} p_i log_3 p_i = -\left[\frac{4}{15} log_3 \frac{4}{15} + \frac{3}{15} log_3 \frac{3}{15} + \frac{8}{15} log_3 \frac{8}{15}\right) = 0.918$
- Based on Profession : 9 Business, 6 Service
- Entropy (Profession) = $\frac{9}{15}\ Entropy(business) + \frac{6}{15}\ Entropy$ (service) = $\frac{9}{15}\left(-\frac{3}{9} log_3 \frac{3}{9} - \frac{6}{9} log_3 \frac{6}{9}\right) + \frac{6}{15}\left(-\frac{1}{6} log_3 \frac{1}{6} - \frac{2}{6} log_3 \frac{2}{6} -\right.$


Credit Rating
(2,5]
(1.9,2]
(1.7,1.9]
(0,1.7]
Buys laptop with CD Writer
Profession (Service)
Buys Laptop with Printer
Buys only laptop
P=0.5
P=0.5
Buys laptop with CD Writer
Buys laptop with printer

The content of the slides are prepared from different textbooks.

References:

- Data Mining and Predictive Analytics, By Daniel T. Larose. Copyright 2015 John Wiley & Sons, Inc.
- Predictive Analytics for Dummies, By Anasse Bari, Mohamed Chaouchi, & Tommy Jung, Copyright 2016, John Wiley & Sons, Inc.
- Introduction to Data Mining with Case Studies, By G.K. Gupta. Copyright 2014 by PHI Learning Private Limited.


Thank you..