Decision Tree

Theme

Data

In[13]:=
```
attributeNames = {"Temperature", "Guests", "Food"};
```

In[14]:=
```
warm = "Warm";
cold = "Cold";
nothing = "Nothing";
snacks = "Snacks";
vegetables = "Vegetables";
flop = "Flop";
hit = "Hit";
label = 4;
```

In[22]:=
```
data = {{cold, 10, nothing, flop},
        {cold, 20, vegetables, hit},
        {cold, 2, vegetables, flop},
        {cold, 8, snacks, hit},
        {warm, 30, snacks, hit},
        {warm, 5, nothing, flop},
        {warm, 28, nothing, hit}};
```

In[23]:=
```
MapThread[Prepend, {data, Range[Dimensions[data, 1][[1]]]}] // TableForm
```

Out[23]//TableForm=

1	Cold	10	Nothing	Flop
2	Cold	20	Vegetables	Hit
3	Cold	2	Vegetables	Flop
4	Cold	8	Snacks	Hit
5	Warm	30	Snacks	Hit
6	Warm	5	Nothing	Flop
7	Warm	28	Nothing	Hit

Level of Measurement

Regarding the level of measurement:

- Here, temperature is used as ordinal-scaled feature (not interval-scaled).
- Number of guests is a ratio scaled feature.
- Food is also a nominal-scaled feature.

Implementation

Result

In[35]:=
```
{tree, nodes, edges} = findTree[];
```

Impurity of parent All: $\frac{3}{7}$

Checking attribute Temperature

Split: $\langle \left| \text{Cold} \to \left\{ \{\{\text{Cold}, 10, \text{Nothing}, \text{Flop}\}, \{\text{Cold}, 20, \text{Vegetables}, \text{Hit}\}, \{\text{Cold}, 2, \text{Vegetables}, \text{Flop}\}, \{\text{Cold}, 8, \text{Snacks}, \text{Hit}\}\}, \left\{\frac{1}{2}, \frac{1}{2}\right\}\right\}, \right.$
$\left. \text{Warm} \to \left\{ \{\{\text{Warm}, 30, \text{Snacks}, \text{Hit}\}, \{\text{Warm}, 5, \text{Nothing}, \text{Flop}\}, \{\text{Warm}, 28, \text{Nothing}, \text{Hit}\}\}, \left\{\frac{2}{3}, \frac{1}{3}\right\}\right\} \right| \rangle$

Gain: 0.

Checking attribute Guests

Split: $\langle \left| < 13.5833 \to \left\{ \{\{\text{Cold}, 10, \text{Nothing}, \text{Flop}\}, \{\text{Cold}, 2, \text{Vegetables}, \text{Flop}\}, \{\text{Cold}, 8, \text{Snacks}, \text{Hit}\}, \{\text{Warm}, 5, \text{Nothing}, \text{Flop}\}\}, \left\{\frac{1}{4}, \frac{3}{4}\right\}\right\}, \right.$
$\left. \geq 13.5833 \to \left\{ \{\{\text{Cold}, 20, \text{Vegetables}, \text{Hit}\}, \{\text{Warm}, 30, \text{Snacks}, \text{Hit}\}, \{\text{Warm}, 28, \text{Nothing}, \text{Hit}\}\}, \{1, 0\}\right\} \right| \rangle$

Gain: 0.285714

Checking attribute Food

Split:
$\langle \left| \text{Nothing} \to \left\{ \{\{\text{Cold}, 10, \text{Nothing}, \text{Flop}\}, \{\text{Warm}, 5, \text{Nothing}, \text{Flop}\}, \{\text{Warm}, 28, \text{Nothing}, \text{Hit}\}\}, \left\{\frac{1}{3}, \frac{2}{3}\right\}\right\}, \right.$
$\text{Vegetables} \to \left\{ \{\{\text{Cold}, 20, \text{Vegetables}, \text{Hit}\}, \{\text{Cold}, 2, \text{Vegetables}, \text{Flop}\}\}, \left\{\frac{1}{2}, \frac{1}{2}\right\}\right\},$
$\left. \text{Snacks} \to \left\{ \{\{\text{Cold}, 8, \text{Snacks}, \text{Hit}\}, \{\text{Warm}, 30, \text{Snacks}, \text{Hit}\}\}, \{1, 0\}\right\} \right| \rangle$

Gain: 0.142857

⇒ Guests gives the best split

Impurity of parent < 13.5833: $\frac{1}{4}$

Checking attribute Temperature

Split: $\langle \left| \text{Cold} \to \left\{ \{\{\text{Cold}, 10, \text{Nothing}, \text{Flop}\}, \{\text{Cold}, 2, \text{Vegetables}, \text{Flop}\}, \{\text{Cold}, 8, \text{Snacks}, \text{Hit}\}\}, \left\{\frac{1}{3}, \frac{2}{3}\right\}\right\}, \right.$
$\left. \text{Warm} \to \left\{ \{\{\text{Warm}, 5, \text{Nothing}, \text{Flop}\}\}, \{0, 1\}\right\} \right| \rangle$

Gain: 0.

Checking attribute Guests

Split: $\langle \left| < 6.83333 \to \left\{ \{\{\text{Cold}, 2, \text{Vegetables}, \text{Flop}\}, \{\text{Warm}, 5, \text{Nothing}, \text{Flop}\}\}, \{0, 1\}\right\}, \right.$
$\left. \geq 6.83333 \to \left\{ \{\{\text{Cold}, 10, \text{Nothing}, \text{Flop}\}, \{\text{Cold}, 8, \text{Snacks}, \text{Hit}\}\}, \left\{\frac{1}{2}, \frac{1}{2}\right\}\right\} \right| \rangle$

Gain: 0.

Checking attribute Food

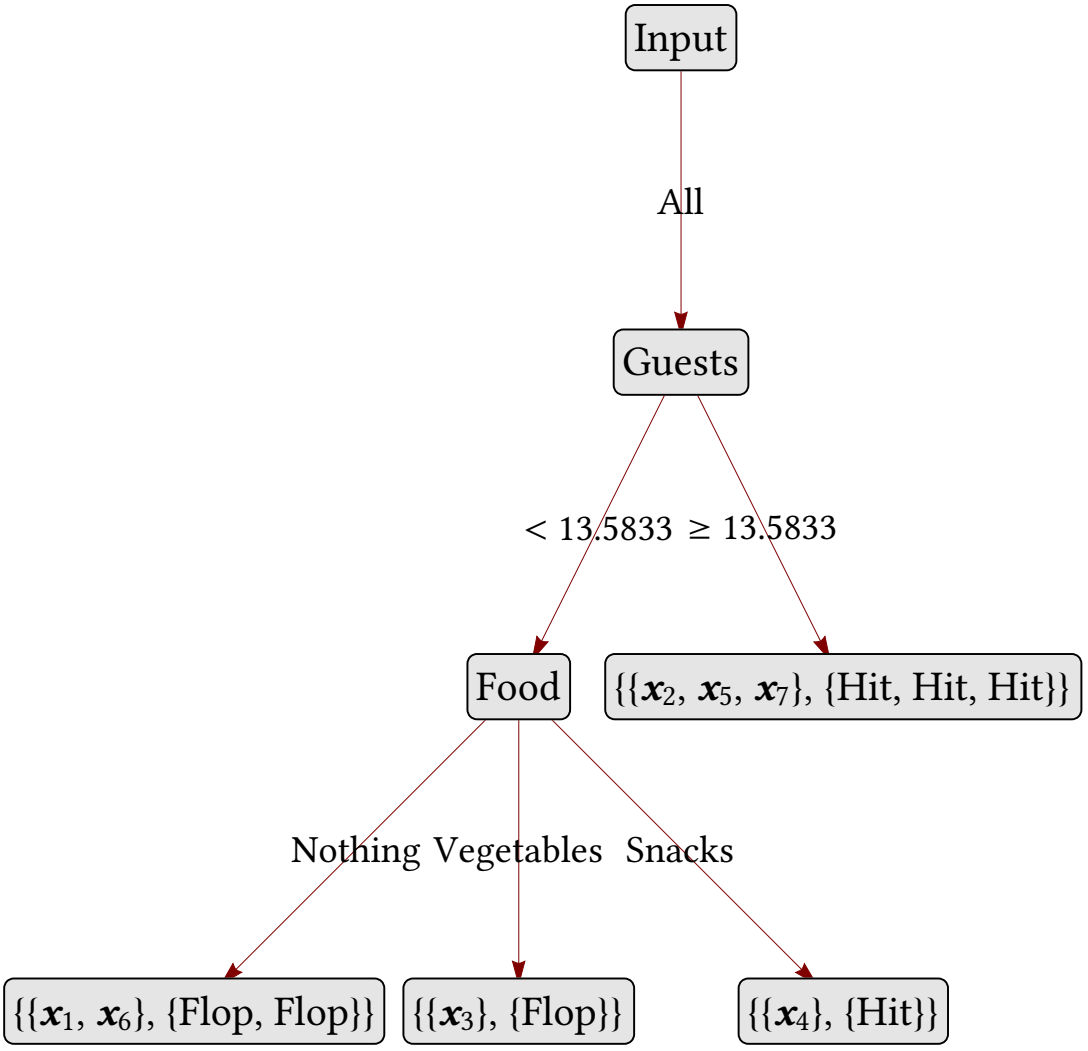
Split: $\langle \left| \text{Nothing} \to \left\{ \{\{\text{Cold}, 10, \text{Nothing}, \text{Flop}\}, \{\text{Warm}, 5, \text{Nothing}, \text{Flop}\}\}, \{0, 1\}\right\}, \right.$
$\left. \text{Vegetables} \to \{\{\{\text{Cold}, 2, \text{Vegetables}, \text{Flop}\}\}, \{0, 1\}\}, \text{Snacks} \to \left\{ \{\{\text{Cold}, 8, \text{Snacks}, \text{Hit}\}\}, \{1, 0\}\right\} \right| \rangle$

Gain: 0.25

⇒ Food gives the best split

In[36]:=
```
TreeGraph[tree,
 GraphLayout → {"LayeredEmbedding", "RootVertex" → 1},
 VertexLabels → Table[
   a → Placed[
     Framed[Style[ToString[nodes[a], StandardForm], FontFamily → "Libertinus Serif",
       FontSize → 22], Background → GrayLevel[0.9], RoundingRadius → 5],
     Center
    ]
   , {a, Keys@nodes}],
 EdgeLabels → Table[
   a → Placed[
     Style[ToString@edges[a], FontFamily → "Libertinus Serif", FontSize → 20],
     Center
    ]
   , {a, Keys@edges}],
 LabelStyle → Directive[FontFamily → "Libertinus Serif", FontSize → 22],
 ImageSize → 600,
 ImagePadding → {{100, 120}, {10, 10}},
 ImageMargins → 0,
 GraphStyle → "VintageDiagram"
]
```

Out[36]=



A note regarding the splitting of the continuous feature: there are several ways of calculating θ. This may be a suitable question for the students in the lab. Some possibilities are:

- We could use a brute-force approach by testing every midpoint between all values and select the one with the highest gain.

- A more sophisticated approach would be to apply SVM on the data and use the decision line (or rather a point in this case) as threshold.