DECISION TREES

Data collection → Data processing → Exploratory analysis & Data viz → Analysis, hypothesis testing, & ML → Insight & Policy Decision
INTRODUCTION

• Let us say, we have two types of vertebrates: mammals and non-mammals

• Suppose a new species is discovered.

• How can we tell whether it is a mammal or a non-mammal.
MAMMAL OR NON-MAMMAL

- Ask a series of questions...

Body temperature?
- Warm
  - Gives birth
    - Yes
      - Mammal
    - No
      - Non-mammal
  - Cold
    - Non-mammal
DECISION TREES

• Root node - no incoming edges and zero or more outgoing edges

• Internal nodes - each has exactly one incoming edge and two or more outgoing edges.

• Leaf or terminal nodes — exactly one incoming edge and no outgoing edges.

• Each leaf node gets a class label.
### Decision Trees - Loan Payment Default?

<table>
<thead>
<tr>
<th>Index</th>
<th>Home Owner</th>
<th>Marital Status</th>
<th>Annual Income</th>
<th>Deafulted borrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125k</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100k</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70k</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120k</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95k</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60k</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220k</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85k</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75k</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90k</td>
<td>Yes</td>
</tr>
</tbody>
</table>
DEFAULTED OR NOT

<table>
<thead>
<tr>
<th>Index</th>
<th>Home Owner</th>
<th>Marital Status</th>
<th>Annual Income</th>
<th>Defaulted borrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125k</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100k</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70k</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120k</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95k</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60k</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220k</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85k</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75k</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90k</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- **Home Owner**: Yes
  - **Defaulted = No**
  - **Annual Income**
    - < 80K: Defaulted = No
    - >= 80K: Defaulted = Yes
- **Home Owner**: No
  - **Marital status**
    - **Single, divorced**
    - **Annually Income**
      - < 80K: Defaulted = No
      - >= 80K: Defaulted = Yes
    - **Married**
      - **Defaulted = No**
DEFAULTED OR NOT

• How do you split?

• When should the splitting procedure stop?
ATTRIBUTE TEST CONDITIONS

• Binary attributes - Two potential outcomes

  Home Owner
  Yes  No

• Nominal attributes

  Marital status
  Single  Divorced  Married

Multi way
ATTRIBUTE TEST CONDITIONS

- Nominal attributes

Marital status
- Single
- Divorced
- Married

Marital status
- Married
- Single, Divorced

Marital status
- Single
- Married, Divorced

Marital status
- Single, Married
- Divorced

Binary split by grouping attribute values
ATTRIBUTE TEST CONDITIONS

• Ordinal attributes

- Shirt size
  - Married, Medium
  - Large, extra large

- Shirt size
  - Small
  - Medium, large, extra large

- Shirt size
  - Small, large
  - Medium, Extra large
ATRIBUTE TEST CONDITIONS

- Continuous attributes

Annual income > 80k

Yes  No

Annual income

<10K  {10k, 25k}  {25K, 50k}  {50K, 80k}  >80K
MEASURES FOR SELECTING THE BEST SPLIT

• Class distribution of records before and after split.
IMPURITY MEASURES

- Selection of the best split based on degree of impurity of the child nodes

- The smaller the degree of impurity, the more skewed the class distribution

- For example, a node with class distribution (0, 1) has zero impurity, whereas a node with uniform distribution (0.5, 0.5) has the highest impurity.
IMPURITY MEASURES

\[ \text{entropy}(t) = - \sum_{i=0}^{c-1} \frac{p(i \mid t) \log_2 p(i \mid t)}{} \]

\[ Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i \mid t)]^2 \]

Classification error(t) = 1 – \( \max_i[p(i \mid t)] \)
CLASSIFICATION ERROR

\[ Error = \frac{\#\text{Incorrect predictions}}{\#\text{Examples}} \]

- Find the decision tree with the least error.

- It is an NP-complete problem
GREEDY ALGORITHM

Step 1: Start with empty tree
Step 2: Select a feature to split data
for each split of the tree:
   Step 3: if nothing more to split, make predictions
   Step 4: else go to step 2 & continue on the split.
## Weather Data

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
</tbody>
</table>
Weather Data

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cloudy</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
</tbody>
</table>

Step 2: Select a feature

Outlook
- sunny
  - yes
  - yes
  - no
  - no

Overcast
- yes
- yes
- yes
- no

Rainy
- yes
- yes
- yes
- yes

Temperature
- hot
  - yes
  - yes
  - yes
  - no
- mild
  - yes
  - yes
  - yes
  - yes
- cool
  - yes
  - yes
  - no
  - no

Humidity
- high
  - yes
  - yes
  - yes
  - no
- normal
  - yes
  - yes
  - yes
  - yes

Windy
- true
  - yes
  - yes
  - yes
  - no
- false
  - yes
  - yes
  - yes
  - no
Measure of purity - information gain

\[
\text{entropy}(t) = - \sum_{i=0}^{c-1} p(i \mid t) \log_2 p(i \mid t)
\]

\[
Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i \mid t)]^2
\]

\[
\text{Classification error}(t) = 1 - \max_i [p(i \mid t)]
\]
Information Gain - outlook

\[ I(\text{sunny}) = - \left( \frac{2}{5} \right) \log_2 \left( \frac{2}{5} \right) - \left( \frac{3}{5} \right) \log_2 \left( \frac{3}{5} \right) \]
\[ = 0.528 + 0.4421 = 0.9701 \]

\[ I(\text{overcast}) = -\left( \frac{4}{4} \right) \log_2 \left( \frac{4}{4} \right) - \left( \frac{0}{4} \right) \log_2 \left( \frac{0}{4} \right) \]
\[ = 0 \]

\[ I(\text{rainy}) = -\left( \frac{3}{5} \right) \log_2 \left( \frac{3}{5} \right) - \left( \frac{2}{5} \right) \log_2 \left( \frac{2}{5} \right) \]
\[ = 0.9701 \]

Weighted average,
\[ I(\text{outlook}) = \left( \frac{5}{14} \right) \times 0.9701 + \left( \frac{4}{14} \right) \times 0 + \left( \frac{5}{14} \right) \times 0.9701 = 0.6929 \]

Information before split,
\[ I([9,5]) = -\left( \frac{9}{14} \right) \log_2 \left( \frac{9}{14} \right) - \left( \frac{5}{14} \right) \log_2 \left( \frac{5}{14} \right) \]
\[ = 0.940286 \]

\[ IG(\text{outlook}) = I([9,5]) - I(\text{outlook}) \]
\[ = 0.940286 - 0.6929 \]
\[ = 0.247386 \]
Information Gain - temperature

\[ I(\text{hot}) = - \frac{2}{4} \log_2 \left( \frac{2}{4} \right) - \frac{2}{4} \log_2 \left( \frac{2}{4} \right) \]
\[ = 0.5 + 0.5 = 1 \]

\[ I(\text{mild}) = - \frac{4}{6} \log_2 \left( \frac{4}{6} \right) - \frac{2}{6} \log_2 \left( \frac{2}{6} \right) \]
\[ = 0.3899 + 0.5283 = 0.9182 \]

\[ I(\text{cool}) = - \frac{3}{4} \log_2 \left( \frac{3}{4} \right) - \frac{1}{4} \log_2 \left( \frac{1}{4} \right) \]
\[ = 0.8112 \]

Weighted average,
\[ I(\text{temperature}) = \frac{4}{14} \times 1 + \frac{6}{14} \times 0.9182 + \frac{4}{14} \times 0.8112 = 0.911 \]

Information before split,
\[ I([9,5]) = - \frac{9}{14} \log_2 \left( \frac{9}{14} \right) - \frac{5}{14} \log_2 \left( \frac{5}{14} \right) \]
\[ = 0.940286 \]

\[ IG(\text{temperature}) = I([9,5]) - I(\text{temperature}) \]
\[ = 0.940286 - 0.911 \]
\[ = 0.02928 \]
Information Gain - humidity

\[ I(\text{normal}) = - \left( \frac{6}{7} \right) \log_2 \left( \frac{6}{7} \right) - \left( \frac{1}{7} \right) \log_2 \left( \frac{1}{7} \right) \]
\[ = 0.19 + 0.401 \]
\[ = 0.591 \]

\[ I(\text{high}) = - \left( \frac{3}{7} \right) \log_2 \left( \frac{3}{7} \right) - \left( \frac{4}{7} \right) \log_2 \left( \frac{4}{7} \right) \]
\[ = 0.524 + 0.4613 \]
\[ = 0.9853 \]

Weighted average,
\[ I(\text{normal, high}) = \left( \frac{7}{14} \right) x 0.591 + \left( \frac{7}{14} \right) x 0.9853 \]
\[ = 0.7881 \]

Information before split,
\[ I([9,5]) = - \left( \frac{9}{14} \right) \log_2 \left( \frac{9}{14} \right) - \left( \frac{5}{14} \right) \log_2 \left( \frac{5}{14} \right) \]
\[ = 0.940286 \]

IG(humidity) = \[ I([9,5]) - I(\text{humidity}) \]
\[ = 0.940286 - 0.7881 \]
\[ = 0.1521 \]
Information Gain - windy

\[
\begin{align*}
I(\text{true}) &= -(3/6) \log_2 (3/6) - (3/6) \log_2 (3/6) \\
&= -(1/2) \log_2 (1/2) - (1/2) \log_2 (1/2) \\
&= 1
\end{align*}
\]

\[
\begin{align*}
I(\text{false}) &= -(6/8) \log_2 (6/8) - (2/8) \log_2 (2/8) \\
&= -(3/4) \log_2 (3/4) - (1/4) \log_2 (1/4) \\
&= 0.8112781
\end{align*}
\]

Weighted average,
\[
\begin{align*}
I(\text{true}, \text{false}) &= (6/14) \times 1 + (8/14) \times 0.8112781 \\
&= 0.8921589
\end{align*}
\]

Information before split,
\[
\begin{align*}
I(\text{windy}) &= -(9/14) \log_2 (9/14) - (5/14) \log_2 (5/14) \\
&= 0.940286
\end{align*}
\]

\[
\begin{align*}
IG(\text{windy}) &= I(\text{windy}) - I(\text{true}, \text{false}) \\
&= 0.940286 - 0.8921589 \\
&= 0.0481271
\end{align*}
\]
Information Gain

- \( \text{gain(} \text{outlook} \text{)} = 0.247 \)
- \( \text{gain(} \text{temperature} \text{)} = 0.029 \)
- \( \text{gain(} \text{humidity} \text{)} = 0.152 \)
- \( \text{gain(} \text{windy} \text{)} = 0.048 \)
Information Gain - Sunny - Temp

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
</tbody>
</table>

\[ I(\text{hot}) = -(2/2) \times \log(2/2) = 0 \]

\[ I(\text{mild}) = -(1/2) \times \log(1/2) = 1 \]

\[ I(\text{cool}) = -1 \times \log(1) = 0 \]

weighted average,

\[ I(\text{temp}) = 0\times(2/5) + 1 \times (2/5) + 0 \times (1/5) = 0.4 \]

\[ I([2,3]) = -(2/5) \times \log(2/5) - 3/5 \times (\log 3/5) = 0.5287 + 0.4421 = 0.971 \]

\[ IG(\text{temperature}) = I([2,3]) - I(\text{temp}) = 0.971 - 0.4 = 0.571 \]
Information Gain- Sunny-humidity

\[ \text{Outlook} \]

- sunny
- overcast
- rainy

\[ \text{humidity} \]

- high
- normal

\[ \text{no} \]

- yes

\[ \text{high} \]

\[ \text{normal} \]

\[ \text{no} \]

\[ \text{yes} \]

\[ I(\text{high}) = -(3/3) \times \log(3/3) = 0 \]

\[ I(\text{normal}) = -(2/2) \times \log(2/2) = 0 \]

weighted average,
\[ I(\text{humidity}) = 0 \times (3/5) + 0 \times (2/5) = 0 \]

\[ I([3,2]) = -(3/5) \log(3/5) - 2/5 \log(2/5) \]
\[ = 0.971 \]

\[ IG(\text{humidity}) = I([2,3]) - I(\text{humidity}) = 0.971 - 0 = 0.971 \]
Information Gain - Sunny-Windy

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
</tbody>
</table>

\[
I(\text{true}) = -\left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) = 1
\]

\[
I(\text{false}) = -\left(\frac{2}{3}\right) \log_2 \left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) \log_2 \left(\frac{1}{3}\right) = 0.3899 + 0.5283 = 0.9182
\]

Weighted average,
\[
I(\text{humidity}) = 0.9182 \times \left(\frac{3}{5}\right) + 1 \times \left(\frac{2}{5}\right) = 0.950
\]

\[
I([3,2]) = -\left(\frac{3}{5}\right) \log_2 \left(\frac{3}{5}\right) - \left(\frac{2}{5}\right) \log_2 \left(\frac{2}{5}\right) = 0.971
\]

IG(humidity) = I([3,2]) - I(windy) = 0.971 - 0.950 = 0.021
Information Gain - Sunny

- IG(Temperature) = 0.571
- IG(humidity) = 0.971
- IG(windy) = 0.02
Information Gain - Sunny

- IG(Temperature) = 0.571
- IG(humidity) = 0.971
- IG(windy) = 0.02
Information Gain - Rainy - Temp

\[
\text{I}(\text{mild}) = -(2/3)x\log(2/3) - (1/3)\log(1/3) = 0.3899
\]

\[
\text{I}(\text{cool}) = -(1/2)\log(1/2) - (1/2)\log(1/2) = 1
\]

weighted average,
\[
\text{I}(\text{temp}) = 0.3899 \times (3/5) + 1 \times (2/5) = 0.6339
\]

\[
\text{I}([2,3]) = -(2/5)\log(2/5) - 3/5(\log 3/5)
\]
\[
= 0.5287 + 0.4421 = 0.971
\]

\[
\text{IG(temperature)} = \text{I}([2,3]) - \text{I}(\text{temp}) = 0.971 - 0.6339 = 0.337
\]
Information Gain - Rainy-humidity

\[
I(\text{high}) = -\left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log\left(\frac{1}{2}\right) = 1
\]

\[
I(\text{normal}) = -\left(\frac{2}{3}\right)x\log\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log\left(\frac{1}{3}\right) = 0.3899
\]

weighted average,
\[
I(\text{humidity}) = 1\times\left(\frac{2}{5}\right) + 0.3899\times\left(\frac{3}{5}\right) = 0.09356
\]

\[
I([3,2]) = -(3/5)\log(3/5) - 2/5(\log2/5) = 0.971
\]

\[
IG(\text{humidity}) = I([2,3]) - I(\text{humidity}) = 0.971 - 0.09356 = 0.87744
\]
Information Gain - Rainy-Windy

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
</tbody>
</table>

\[
I(\text{true}) = -(2/2) \times \log(2/2) = 0
\]

\[
I(\text{false}) = -(3/3) \times \log(3/3) = 0
\]

weighted average,

\[
I(\text{humidity}) = 0 \times (3/5) + 0 \times (2/5) = 0
\]

\[
I([3,2]) = -(3/5) \log(3/5) - 2/5 \log(2/5) = 0.971
\]

\[
IG(\text{humidity}) = I([2,3]) - I(\text{windy}) = 0.971 - 0 = 0.971
\]
Information Gain - Rainy

- IG(Temperature) = 0.337
- IG(humidity) = 0.877
- IG(windy) = 0.971

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>cold</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Sunny</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>Overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>Sunny</td>
<td>cool</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
</tbody>
</table>

The decision tree for Rainy weather based on Outlook, Humidity, and Windy conditions.