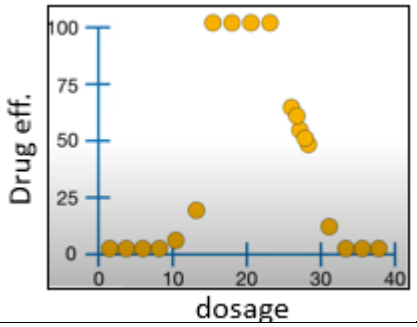
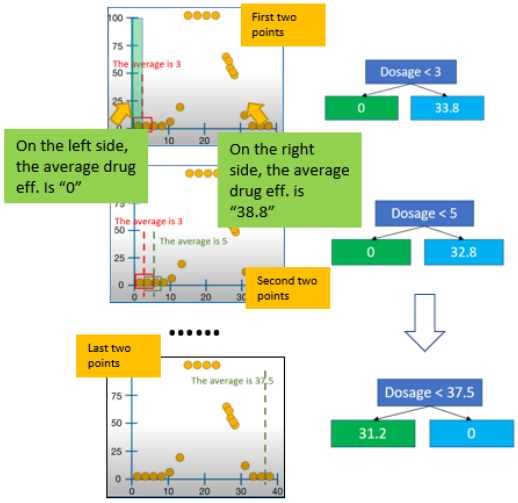
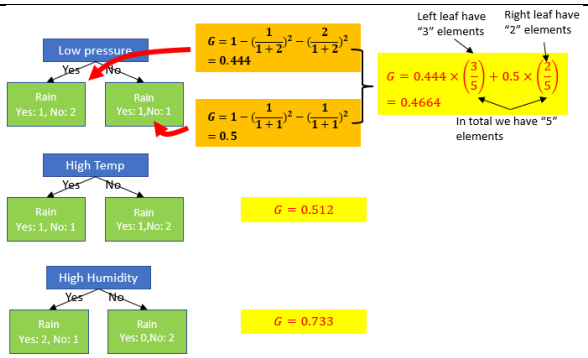
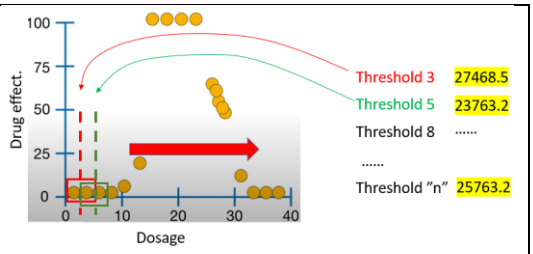
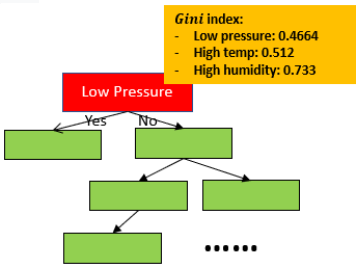
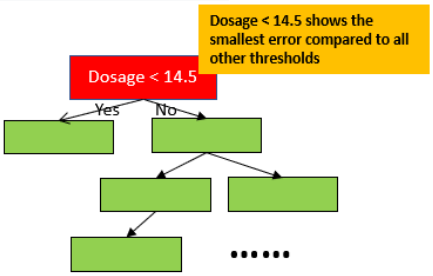
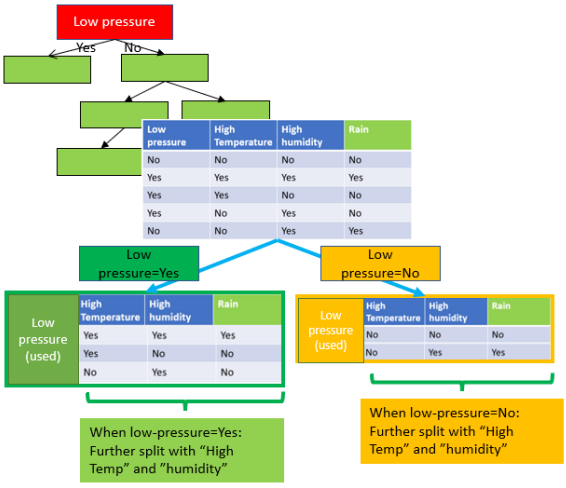
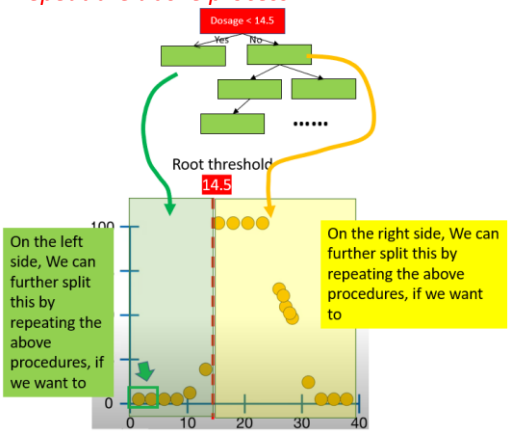
Summary: Decision Tree

<table>
<tr><td></td><td>Classification</td><td>Regression</td></tr>
<tr><td>Sample data</td><td>
<table>
<tr><th>Low pressure</th><th>High Temperature</th><th>High humidity</th><th>Rain</th></tr>
<tr><td>No</td><td>No</td><td>No</td><td>No</td></tr>
<tr><td>Yes</td><td>Yes</td><td>Yes</td><td>Yes</td></tr>
<tr><td>Yes</td><td>Yes</td><td>No</td><td>No</td></tr>
<tr><td>Yes</td><td>No</td><td>Yes</td><td>No</td></tr>
<tr><td>No</td><td>No</td><td>Yes</td><td>Yes</td></tr>
</table>
</td><td>
<table>
<tr><th>Dosage</th><th>Drug effect.</th></tr>
<tr><td>10</td><td>58</td></tr>
<tr><td>20</td><td>60</td></tr>
<tr><td>35</td><td>57</td></tr>
<tr><td>5</td><td>44</td></tr>
<tr><td>…</td><td>…</td></tr>
</table>
</td></tr>
<tr><td>1</td><td></td><td>Sort the predictors from small values to big ones, and plot it out:
 
</td></tr>
<tr><td>2</td><td>Grow trees for each predictor individually
e.g., when we have “Low” pressure, there are:
• 1 case has rain,
• 2 cases do not have rain
 
</td><td>Grow trees for each threshold of every predictor:
the threshold is the “mean” value for neighbouring points
 
</td></tr>
<tr><td>3</td><td>Obtain the “Gini” index for each tree
The “Gini” index can be calculated as
$G = 1 - {P_{yes}}^2 - {P_{no}}^2$
Where:
• P_{yes} is the probability of “yes” in a leaf
• P_{no} is the probability of “no” in a leaf</td><td>Obtain the error for each tree (or threshold)
Using the tree (threshold based) to produce the prediction and calculate the error as
$(Pred - Actual)^2_{point1} + (Pred - Actual)^2_{point2} + \cdots$</td></tr>
</table>

<table>
<tr>
<td></td>
<td> 


</td>
</tr>
<tr>
<td>4</td>
<td>Based on the Gini index, determine the "root" to split the tree
Usually, the predictor showing the smallest Gini is used as "root"
 
</td>
<td>Based on the Error, determine the "root" to split the tree
Usually, the threshold showing the smallest error is used as "root"
 
</td>
</tr>
<tr>
<td>5</td>
<td>Further split the tree using the similar method, until we are not able to split
- Grow tree for every remained predictor
- Calculate "Gini" for each tree
- Get the tree with the smallest Gini
- Repeat the above process …
 

……</td>
<td>Further split the tree using the similar method, until we are not able to split
- Calculate the threshold from neighbouring points, and grow trees based on thresholds
- Error for each tree
- Get the tree with the smallest error
- Repeat the above process …
 
</td>
</tr>
<tr>
<td colspan="3">Note:
- In order to avoid overfitting, we may limit the number of levels of the tree to grow</td>
</tr>
</table>