Homework #3

Vinayak Raghupathy
NYUID-N11568565

0. Explain in your own words the difference between Data Classification & Data Clustering.

Data Clustering is an unsupervised learning technique used to group similar instances on the basis of features. Clustering can be achieved by various algorithms like K-means, Hierarchical Clustering etc.

Data Classification is supervised learning technique used to assign predefined tags to instances on the basis of features. K-Nearest Neighbour algorithm & decision tree algorithms are the most famous classification algorithm used in data mining.

Classification algorithm requires training data, classification model is created from training data, then the model is used to classify new instances. Clustering does not require training data & does not assign predefined label to each & every group.

In Clustering statistical concepts are used & datasets are split into subsets with similar features.
Classification uses algorithms to categorize new data according to observation of training set.

Ans. 1 Removed W2 & W4 from the dataset in question so the following dataset is considered on which ID3 is applied

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Windy	Yes	Rich	Cinema
W3	Rainy	No	Rich	Stay in
W4	Sunny	Yes	Poor	Tennis
W5	Windy	Yes	Poor	Cinema
W6	Windy	No	Rich	Shopping
W7	Rainy	Yes	Rich	Stay in
W8	Sunny	No	Rich	Tennis

$$\text{Entropy(Decision)} = \text{Entropy}\left(\frac{3}{8}, \frac{2}{8}, \frac{2}{8}, \frac{1}{8}\right)$$

$$= -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{1}{8}\log_2\frac{1}{8}$$

$$= -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{4}{8}\log_2\frac{2}{8} - \frac{1}{8}\log_2\frac{1}{8}$$

$$= 1.906$$

$$\text{Expected (conditional Entropy(Weather)} = \left(\frac{3}{8}\right)\text{Entropy}\left(\frac{1}{3}, \frac{2}{3}\right)_{\wedge}^{,90} + \frac{3}{8}\,\text{Entropy}\left(\frac{2}{3}, \frac{1}{3}\right)_{\wedge}^{,0,0} + \frac{2}{8}\,\text{Entropy}(1,0,0,0)$$

$$= \frac{3}{8}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{3}{8}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{2}{8} \times \left(-\frac{2}{2}\log_2\frac{2}{2}\right)$$

$$= \frac{3}{8}\left(-\frac{4}{3}\log_2\frac{2}{3} - \frac{2}{3}\log_2\frac{1}{3}\right) + 0$$

$$= 0.689$$

Informational Gain (Weather)

$= 1.0906 - 0.689 = 1.217$

Expected Conditional Entropy (Parents)

$= \frac{5}{8}\text{Entropy}\left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}, 0\right) + \frac{3}{8}\text{Entropy}\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3},$

$= \frac{5}{8}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{1}{5}\log_2\frac{1}{5} - \frac{1}{5}\log_2\frac{1}{5}\right) + \frac{3}{8}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$

$= \frac{5}{8}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{1}{5}\right) + \frac{3}{8}\left(-\frac{3}{3}\log_2\frac{1}{3}\right)$

$= 1.0451$

Informational Gain (Parents)

$= 1.0906 - 1.0451$

$= 0.0455$

Expected Conditional Entropy (Money)

$= \frac{6}{8}\text{Entropy}\left(\frac{2}{6}, \frac{2}{6}, \frac{1}{6}, \frac{1}{6}\right) + \frac{2}{8}\text{Entropy}\left(\frac{1}{2}, \frac{1}{2}, 0, 0\right)$

$= \frac{6}{8} \times \left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6}\right) + \frac{2}{8} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right)$

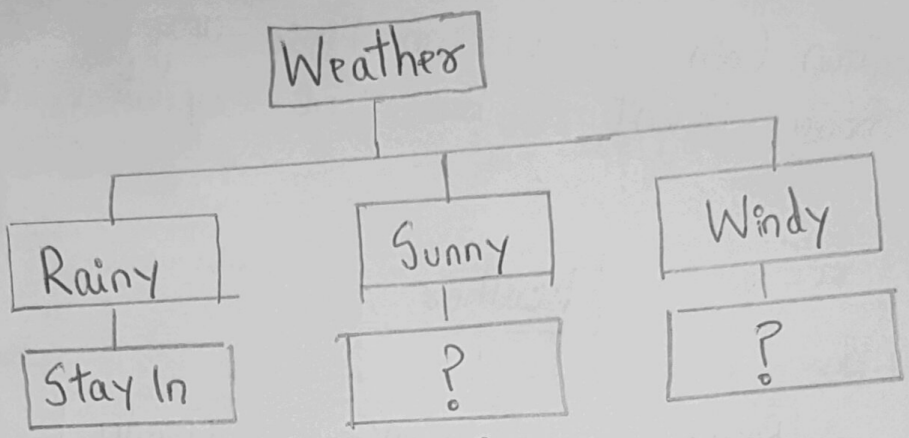
$= \frac{6}{8} \times \left(-\frac{4}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{1}{6}\right) + \frac{2}{8} \times \left(-\frac{2}{2}\log_2\frac{1}{2}\right)$

$= 1.0688$

Informational Gain (Money) $= 1.0906 - 1.0688 = 0.0217$

Choose Weather to be root node since it has highest Information Gain

Updated Decision Tree



Updated Data given Weather is Sunny

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W4	Sunny	Yes	Poor	Tennis
W8	Sunny	No	Rich	Tennis

$$\text{Entropy(Decision)} = \text{Entropy}\left(\frac{1}{3}, \frac{2}{3}\right)$$
$$= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$$

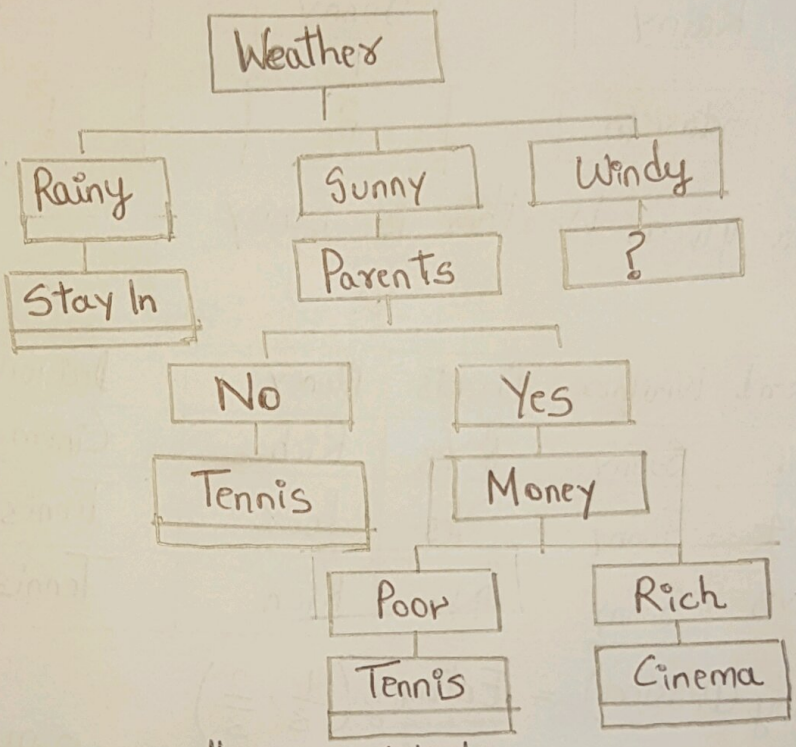
$$\text{Expected Conditional Entropy(Parents)} = \frac{2}{3}\text{Entropy}\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{3}\text{Entropy}(1,0)$$
$$= \frac{2}{3}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + 0$$
$$= 0.667$$

$$\text{Informational Gain(Parents)} = 0.918 - 0.667$$
$$= 0.251$$

$$\text{Expected Conditional Entropy(Money)} = \frac{2}{3}\text{Entropy}\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{3}\text{Entropy}(1,0)$$
$$= \frac{2}{3}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + 0$$
$$= 0.667$$

$$\text{Informational Gain(Money)} = 0.918 - 0.667$$
$$= 0.251$$

We can observe here that attribute Parents & Money both have same Information Gain, so we can choose any of them at first. Suppose we choose Parent then the updated decision tree will be.



Updated data given Weather is Windy

Weekend	Weather	Parents	Money	Decision
W2	Windy	Yes	Rich	Cinema
W5	Windy	Yes	Poor	Cinema
W6	Windy	No	Rich	Shopping

$$\text{Entropy(Decision)} = \text{Entropy}\left(\tfrac{1}{3}, \tfrac{2}{3}\right)$$
$$= -\tfrac{1}{3}\log_2\tfrac{1}{3} - \tfrac{2}{3}\log_2\tfrac{2}{3} = 0.918$$

$$\text{Expected Conditional Entropy (Money)} = \tfrac{2}{3}\text{Entropy}\left(\tfrac{1}{2}, \tfrac{1}{2}\right) + \tfrac{1}{3}\text{Entropy}(1,0)$$
$$= \tfrac{2}{3}\left(-\tfrac{1}{2}\log_2\tfrac{1}{2} - \tfrac{1}{2}\log_2\tfrac{1}{2}\right) + 0 = 0.667$$

Informational Gain (Money) = 0.918 - 0.667

= 0.251

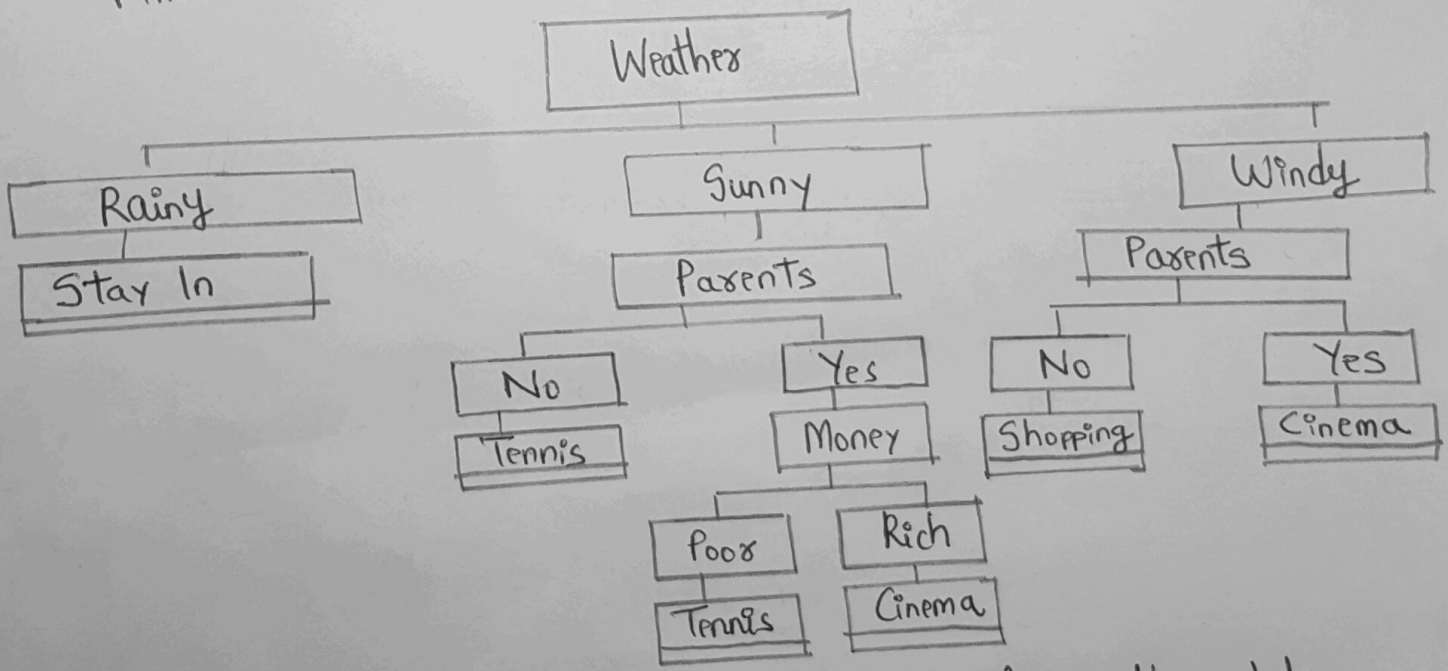
Expected Conditional Entropy (Parents) = $\frac{2}{3}$ Entropy (1, 0) + $\frac{1}{3}$ Entropy (1, 0)

= 0

Informational Gain (Parents) = 0.918

We can observe here Parents will have pure subsets with entropy 0 i.e lowest entropy which would mean highest Informational Gain, so we choose Parents to be our attribute given Weather is Windy

Final Decision Tree



Ans.2 If we use the decision tree above so for the data points

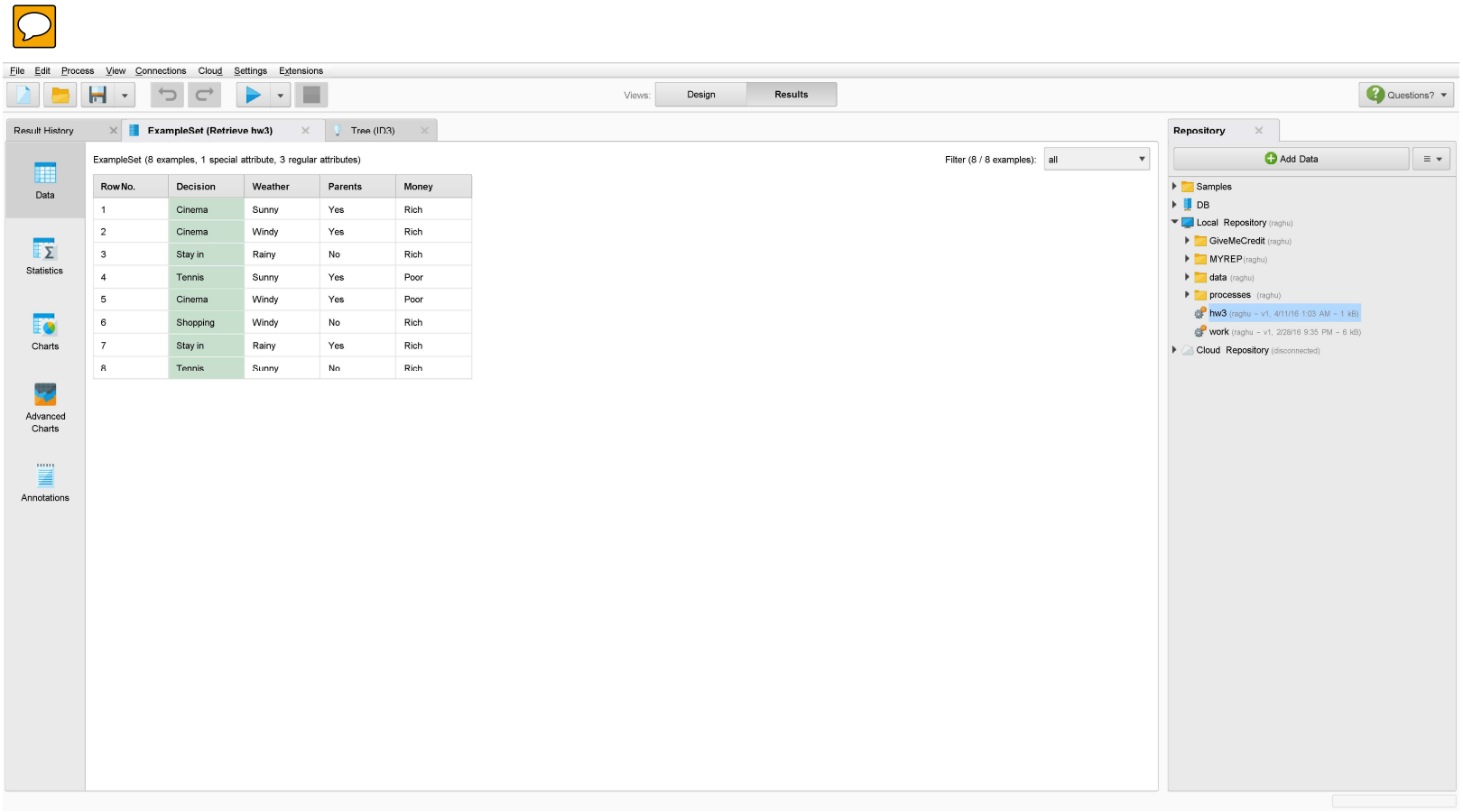
Sunny, Yes, Poor

The decision will be Tennis as the path will go from weather = Sunny & then the value of attribute parents i.e Yes & depending on money attribute we have the outcome as Tennis as Money is Poor

Windy, Yes, Rich

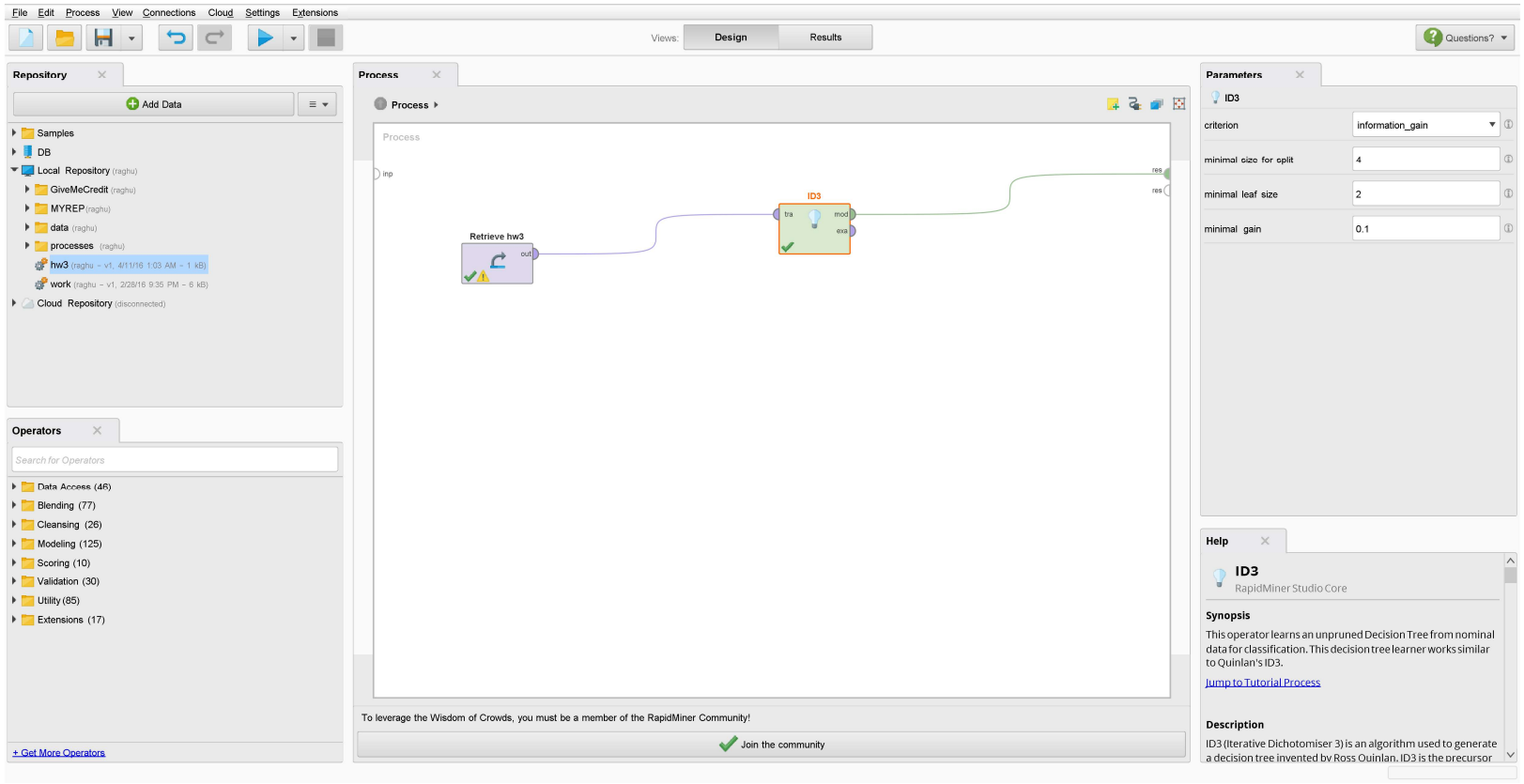
The decision is Cinema as the path will go from weather = windy & reach attribute parent whose value will determine the outcome as its value is yes, so the decision is cinema irrespective of the attribute money's value which is rich here.

Results

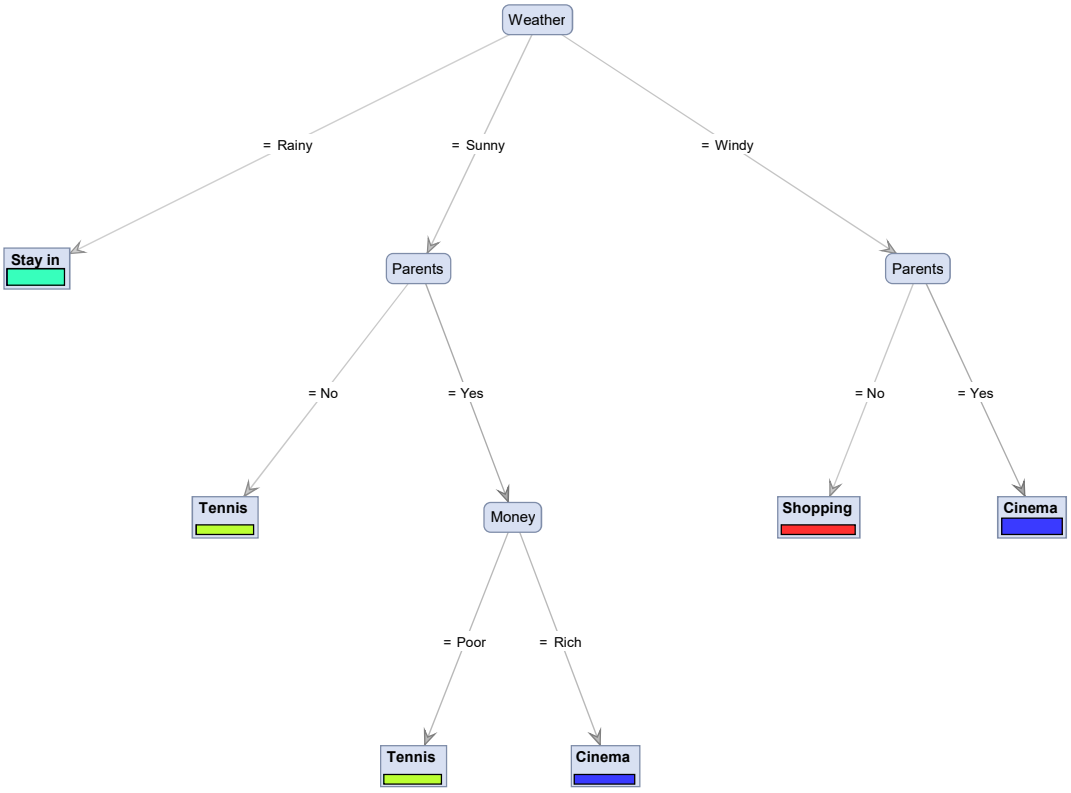


Row No.	Decision	Weather	Parents	Money
1	Cinema	Sunny	Yes	Rich
2	Cinema	Windy	Yes	Rich
3	Stay in	Rainy	No	Rich
4	Tennis	Sunny	Yes	Poor
5	Cinema	Windy	Yes	Poor
6	Shopping	Windy	No	Rich
7	Stay in	Rainy	Yes	Rich
8	Tennis	Sunny	No	Rich

Design


File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
Samples
DB
Local Repository (raghu)
GiveMeCredit (raghu)
MYREP (raghu)
data (raghu)
processes (raghu)
hw3 (raghu – v1, 4/11/16 1:03 AM – 1 kB)
work (raghu – v1, 2/28/16 9:35 PM – 6 kB)
Cloud Repository (disconnected)
Operators
Search for Operators
Data Access (46)
Blending (77)
Cleansing (26)
Modeling (125)
Scoring (10)
Validation (30)
Utility (85)
Extensions (17)
+ Get More Operators
Process
Process
Process
inp
Retrieve hw3
out
ID3
tra
mod
exa
res
res
To leverage the Wisdom of Crowds, you must be a member of the RapidMiner Community!
Join the community
Parameters
ID3
criterion
information_gain
minimal size for split
4
minimal leaf size
2
minimal gain
0.1
Help
ID3
RapidMiner Studio Core
Synopsis
This operator learns an unpruned Decision Tree from nominal data for classification. This decision tree learner works similar to Quinlan's ID3.
Jump to Tutorial Process
Description
ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree invented by Ross Quinlan. ID3 is the precursor

Graph


Weather
= Rainy
= Sunny
= Windy
Stay in
Parents
Parents
= No
= Yes
= No
= Yes
Tennis
Money
Shopping
Cinema
= Poor
= Rich
Tennis
Cinema