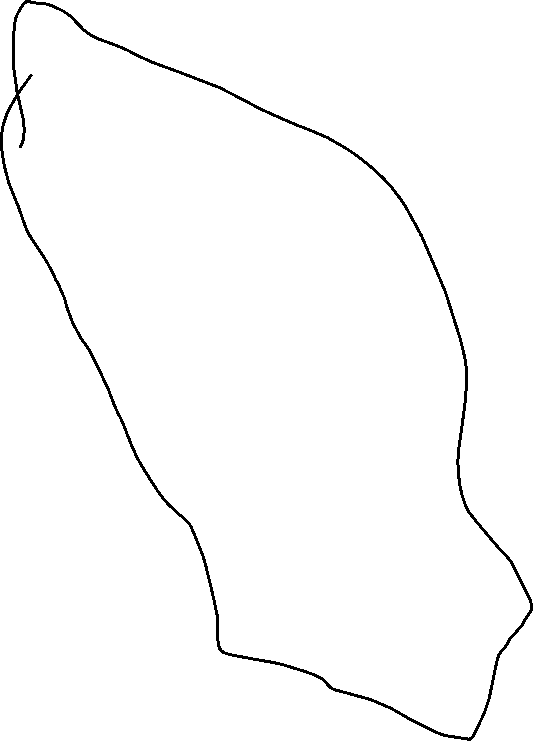
Buy

YES NO

YES NO NO YES



KM Drawn < 50000 km

Don’t Buy

Make = Ferrari

Yes No YES NO

Don’t Buy

Buy

No

Buy

Decision Tree: defining data into nodes of a tree like structure:

* How will you decide which node will come first?
* How will you decide which node will come second and so on and last?
* How will you decide what threshold to be used at a continuous column value: (eg: km>50000)
* How will you decide when to end your tree node?

To answer this question:

* Entropy
* Information gain
* Gini index or Coefficient

Entropy:

E(Subset): -p(+) \* logp (yes) \* -p\*logp (negative)(-)

p- : probability of negative class

p+: probability of positive class

The uncertainty in our dataset or measure of disorder:

Which of the sequence is having more information gain: (more correct predictions)

* Entropy value for full dataset 0.99 after taking a subset of the dataset, information gain reduce by 0.37 if I will make color as red node.

**Minimise information gain:**

**The node which will give you lowest information gain, we will select that as a parent node.**

**Also we need to answer how to decide max level of depth of tree and max feature:**

**Hyperparameter in decision tree:**

**Prunning: pre pruning or Post pruning**

**Python library will consider all our information gain issue and pruning issue on its own.**