# 00:00 Review

# 02:09 TwoR model

OneR model involves developing a model that gives the result using just one binary decision. What is the best result that the model could produce using a single binary decision. In many cases of simple datasets, this is proven to be very effective rather than complicated models. TwoR model is with two consecutive binary splits.

# 04:43 How to create a decision tree

# 07:02 Gini

This is a value that represents the probability of taking two objects from a cluster and the second object being the same category(classification or regression category) as the first one. This ensures that in the clusters, the data points are homogenous and different clusters have different kinds of data points with different values.
Gini value = 0 if we are absolutely sure that the second object is of the same o/p category as the first one. Gini index =1 if the group/cluster is not at all homogenous. So, ideally, the gini index has to be close to zero.

# 10:54 Making a submission

# 15:52 Bagging

# 19:06 Random forest introduction

# 20:09 Creating a random forest

# 22:38 Feature importance

# 26:37 Adding trees

# 29:32 What is OOB

# 32:08 Model interpretation

For tabular data, model interpretation is particularly important. For a given model, the things we are most likely to be interested in are:

- How confident are we in our predictions using a particular row of data?
- For predicting with a particular row of data, what were the most important factors, and how did they influence that prediction?
- Which columns are the strongest predictors, which can we ignore?
- Which columns are effectively redundant with each other, for purposes of prediction?
- How do predictions vary, as we vary these columns?

1) Tree variance for prediction confidence
2)
3) Feature importance plot
4) Check in book-Dendogram
5) Partial dependence plots

# 35:47 Removing the redundant features

# 35:59 What does Partial dependence do

# 39:22 Can you explain why a particular prediction is made

Tree interpreter - In each tree what is the exact path (feature importance plot just for one row)
Using waterfall charts

```
In [82]:  ▶ waterfall(valid_xs_final.columns, contributions[0], threshold=0.08,
                  rotation_value=45,formatting='{:,.3f}');
```
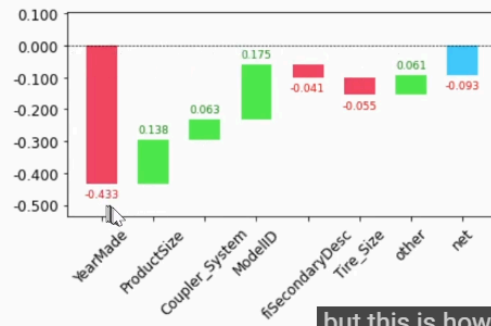


# 46:07 Can you overfit a random forest

# 49:03 What is gradient boosting

Prediction on a small tree, another small tree predicting the error of the first tree and so on
Accuracy compared to random forest
Overfitting?

explained.ai

## 1:32:01 What does our data augmentation do to images

## 1:34:12 How to iterate the model with larger images

## 1:36:08 pandas indexing

## 1:38:16 What data-augmentation does tta use?

Test time augmentation -> Uses the same data augmentation techniques that are used on the training data and then average the results like bagging

Why using image random resize cropping, squishing,padding -> our input image and output image is a square or rectangle? How do they affect our model?

If we observe that a lot of the images in the dataset are of different sizes like more height, less wide and the contrary more wide and less height, it doesn't make sense to make every image into a rectangle using padding as a significant amount of images would get contaminated. Maybe making such a dataset into a square would be a much more suitable option. Hence, it is important for us to test the model accuracy with different kinds of augmentation techniques to see what works best for the specific model and use case. Mostly this chucking into squares is used.