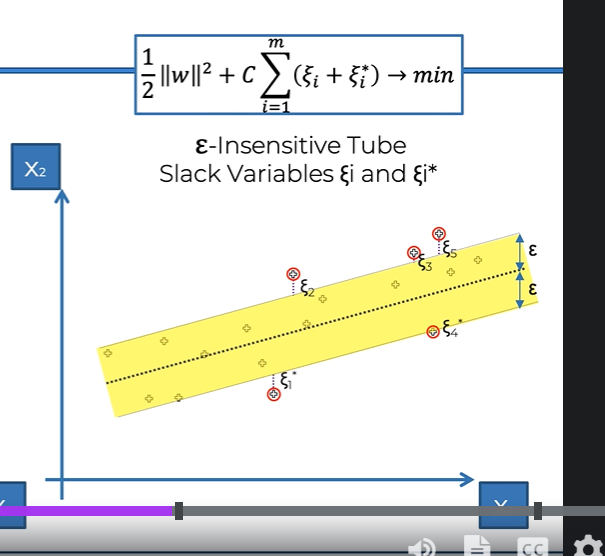
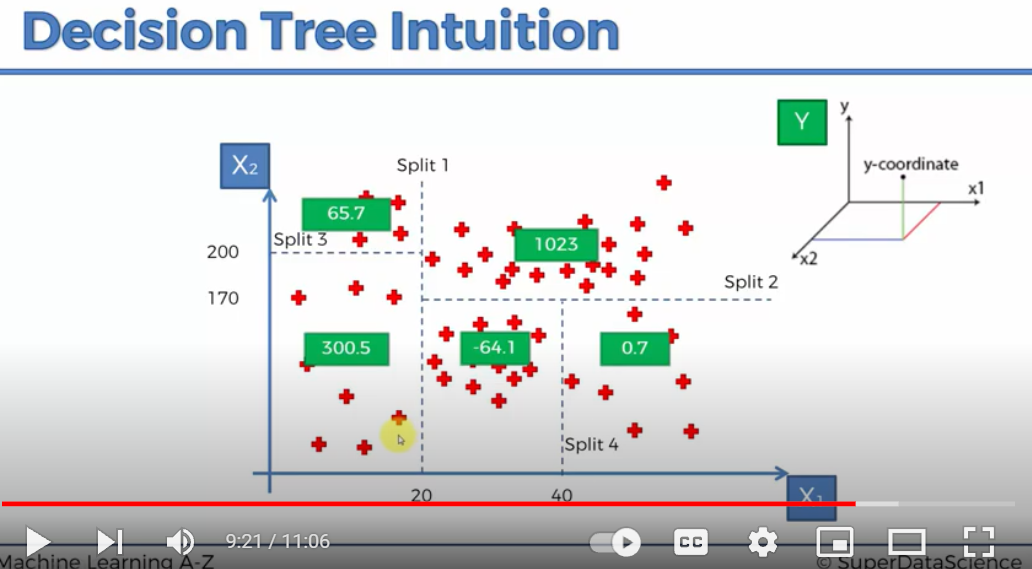
Feature-> Input (Independent variable) , used to predict the attribute: output (dependent variable)

Steps in **preprocessing**->

* Import libraries,
* Import the dataset,
* take care of missing data: Mean, median, dropping data, linear interpolation , most frequent, constant
* Encoding Categorial Data: Label encoding if there is an order (0 for aa and 1 for bb,2 for cc ) and OneHotEncoding if there is no order ( 0 0 1 for A 010 for B and 001 for C)
* Feature Scaling: Standardizing or normalizing the data so different columns are “in the same scale” Standardization: x=x-mean(x)/SD(x) Normalization =x-min(x)/(max(x)-min(x)), Standardization always works

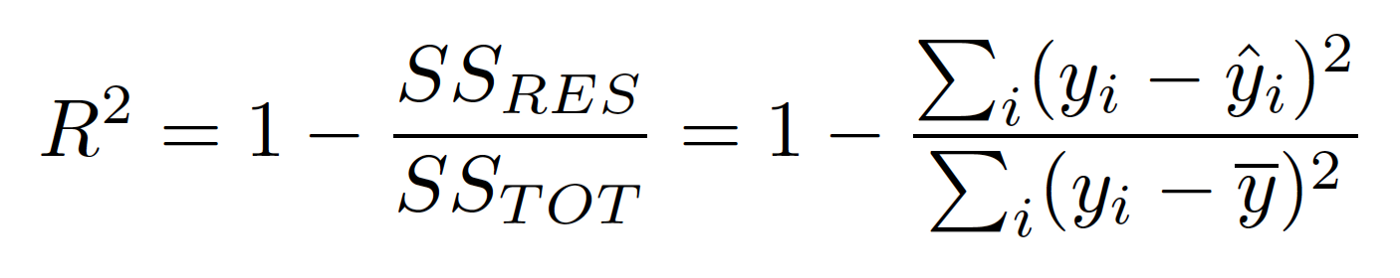
When to apply feature scaling?Ans: When there is no explicit equation between y and x

Regression

1. Simple Linear Regression- > y=mx+b where x is the is the feature and y the output.Single feature used to fit a line through the output
2. Multiple Regression-> y=mx1+m2x2+m3x3, where x1, x2, x3 are the 3 features to predict single output y
3. Polynomial Regression-> y=mx1+m2x1^2+m3x1^3….. with multiple powers of same feature
4. SVR-> A tube of width epsilon ε within which if the points lie, the error is not considered. If points lie outside (slack variables) or support vectors (vectors that dictate how the tube is made). Error is calculated using distance between support vectors and the tube
5. Decision Trees: Split is made in the dependent variables say X1 and X2 based on several conditions like X1>20, X2>50 based on information entropy or when there is a certain minimum number of leaves in that split. A yes no based decision tree is made based on conditions. Once that group of leaves found (terminal leaves), take average of y for all of the points in the terminal leaves to find the output. 
6. Random forest Regressor: Ensemble learning-> Performing the same algorithm multiple times to make much more powerful model. Choose at random K datapoints -> Make a decision tree based on the k datapoints -> Repeat first 2 steps for N number of trees-> Take the average of the output of each tree

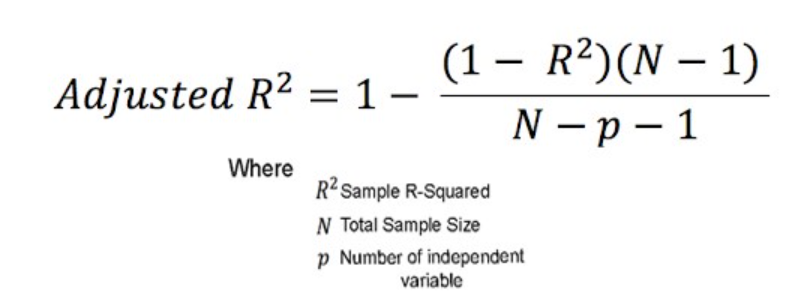
R squared

How good the Residual Sum of squares or the fitted data (datapoint-fitted) is compared to the Total sum of squared (datapoint – average) ( Average line). Higher the SSres, more the R squareds

How much better is the fitted data to the average of data, 

Adjusted R squared:

If new variables added , at the worst, R squared will remain the same or increase. Because there is always some correlation between the extra variable and the output variable, ie R2 is always biased.

R squared-> penalization factor for variables that do not help the E squared to increased.\ 

So if an extra independent variable is added, There is a penalty ass N-p-1 decreases and adj R^2 decreases. But 1-R^2 decreases so adjust R^2 increases. Hence the playoff