# Table of Contents

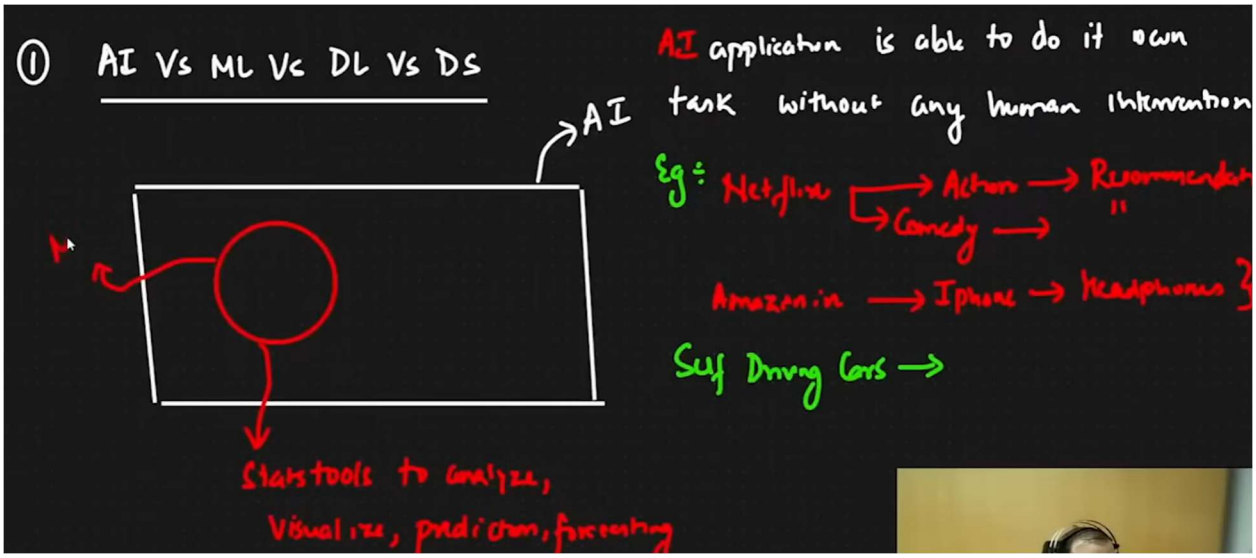


| Supervised Learning Algorithms | Unsupervised Learning Algorithms |
|---|---|
| 1. Linear Regression | 1. K-Means |
| 2. Ridge & Lasso | 2. DBSCAN |
| 3. Logistic Regression | 3. Hierarchical Clustering |
| 4. Decision Tree | 4. K-Nearest Neighbor (Clustering) |
| 5. AdaBoost | 5. PCA (Principal Component Analysis) |
| 6. Random Forest | 6. LDA (Linear Discriminant Analysis) |
| 7. Gradient Boosting | |
| 8. XGBoost | |
| 9. Naive Bayes | |

## AI VS ML VS DL

Machine learning is basically subset of AI

- ML basically gives you stats tool to analyze visualizing , predicting and forcasting




- The goal is to create an AI application

MACHINE LEARNING AND DEEP LEARNING

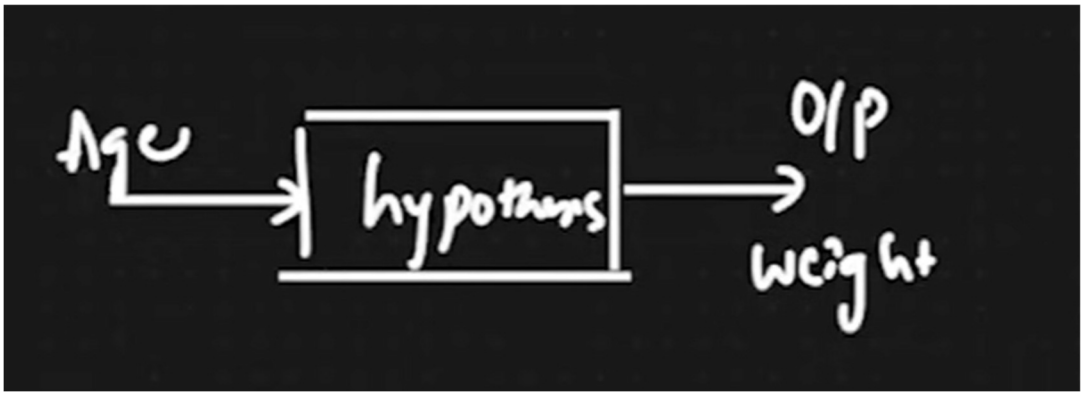
Two algorithm
Supervised ML
Unsupervised ML

| Supervised ML | Unsupervised ML |
|---|---|
| • Regression Problem<br>• Classification Problem | • Clustering<br>• Dimensionality reduction |

Note: There is also one more time which is called reinforcement learning

Supervised ML Example of a data set :

| Age | Weight |
|---|---|
| 24 | 62 |
| 25 | 63 |
| 21 | 72 |
| 27 | 62 |



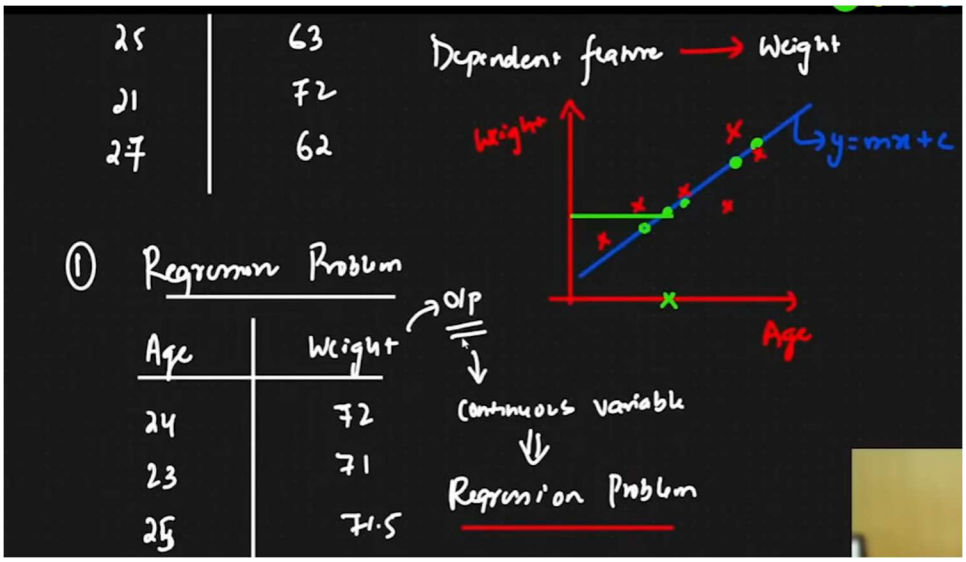

- INDEPENDENT FEATURE (
  - AGE

- DEPENDENT GEATURE
  - Weight

  The value of weight is changing according to age so that's why they are categorized as dependent and independent

  Regression vs Classification

  Regression




- In regression there will be outputs of continuous variable

CLASSIFICATION

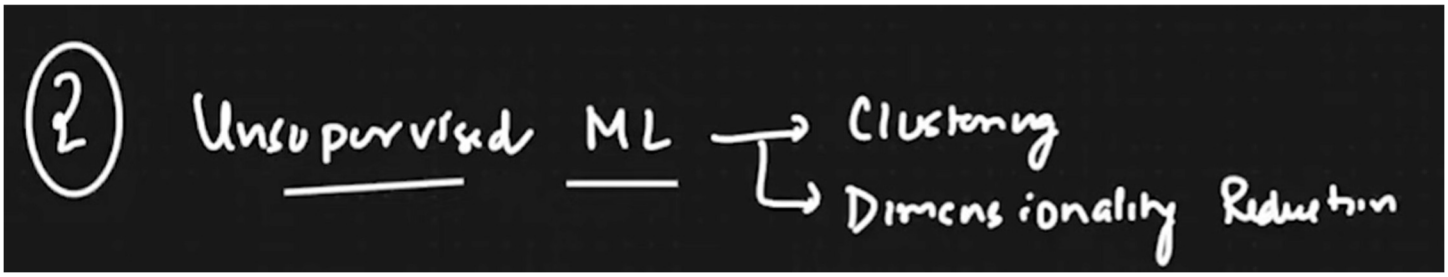
| No of hrs | No of play hrs | No of sleep | Pass or Fail(dependent geature) AKA Putput |
|---|---|---|---|
| - | - | - | P |
| - | - | - | F |

Classification : whenever you have fixed number of categorical output it becomes a classification output

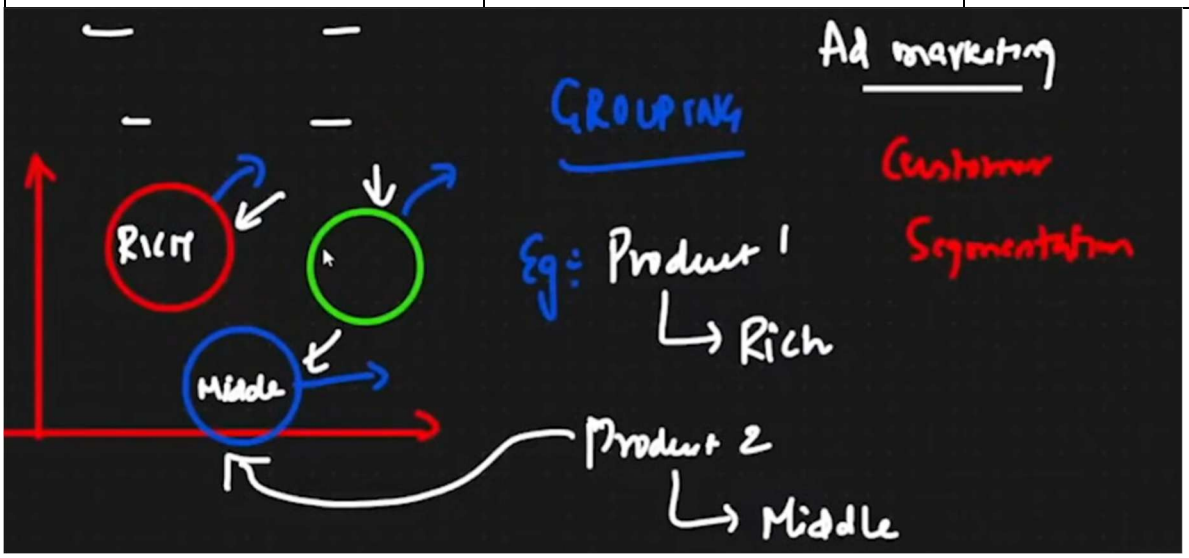
Binary Classification: There are two categories of output Example P/F

Multi class classification: Multiple categories of output

Unsupervised ML




| Salary | Age | No dependent variable |
|---|---|---|
| - | - | |
| - | - | |




- Clustering
    - Grouping
        - Example for Ad marketing : Product 1 for rich people product 2 for poor people etc

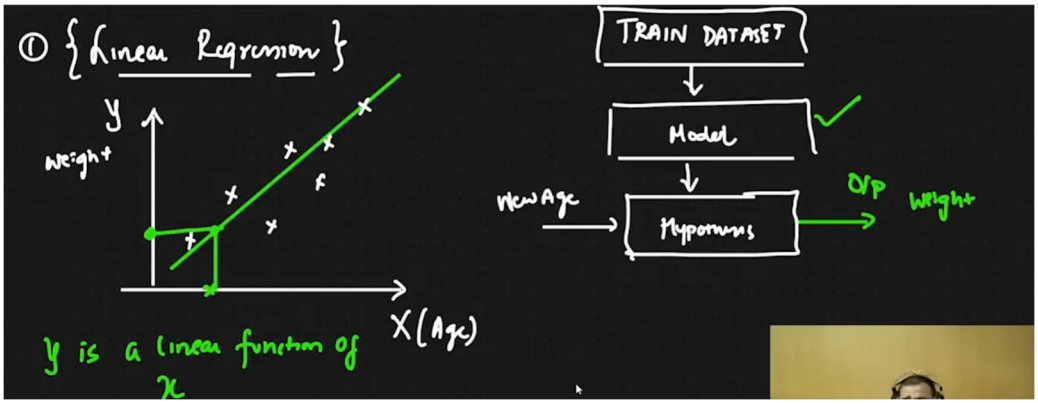
Dimensionality Reduction

Lower 100 feature of 100 feature examples of that is PCA and LDA

| Supervised Learning Algorithms | Unsupervised Learning Algorithms |
|---|---|
| 1. Linear Regression | 1. K-Means |
| 2. Ridge & Lasso | 2. DBSCAN |
| 3. Logistic Regression | 3. Hierarchical Clustering |
| 4. Decision Tree | 4. K-Nearest Neighbor (Clustering) |
| 5. AdaBoost | 5. PCA (Principal Component Analysis) |
| 6. Random Forest | 6. LDA (Linear Discriminant Analysis) |
| 7. Gradient Boosting | |
| 8. XGBoost | |
| 9. Naive Bayes | |

**Linear Regression**

Find out the best fit line which will actually help in making prediction




Y is a linear function of X

The line equation could be given as:

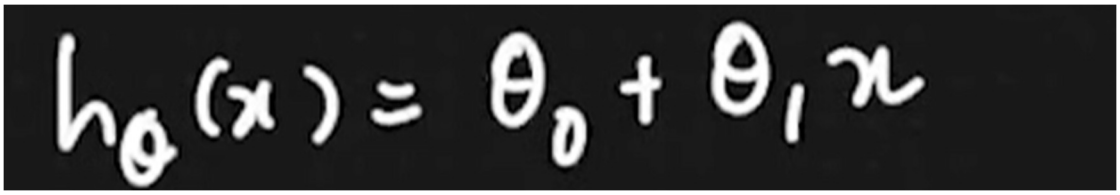
Y=mx+c

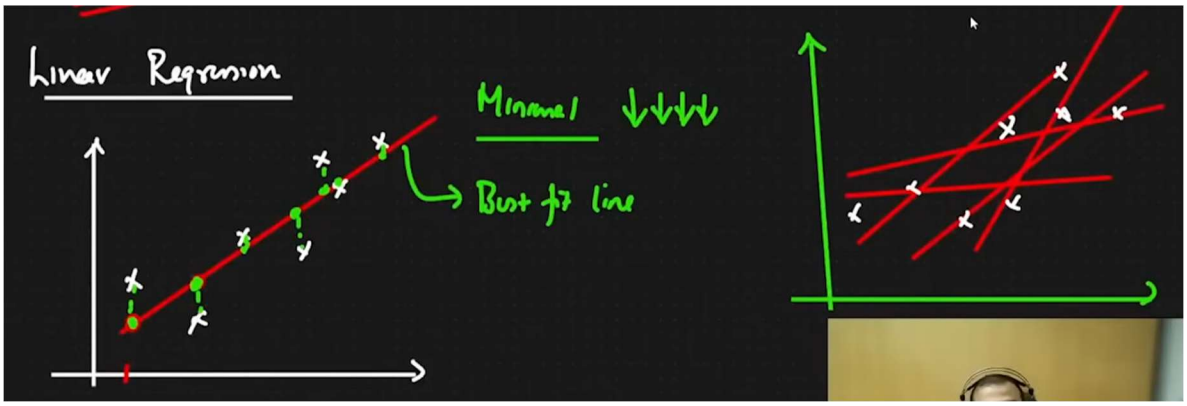
Y=B0+B1x

H(x) = theta0 +theta1x

Etc

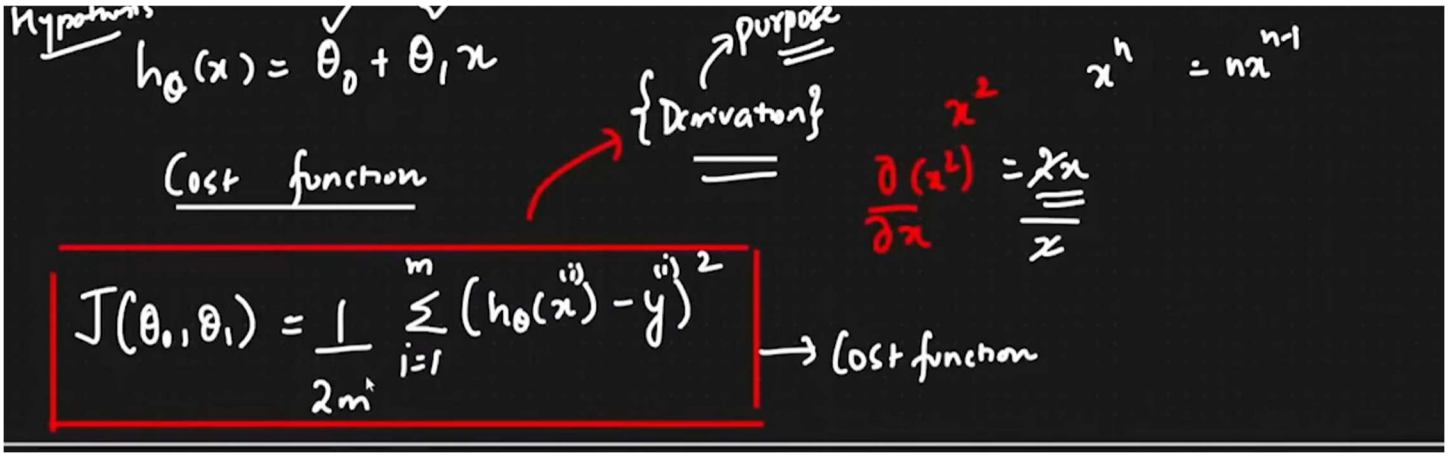
This is the best fit line eqn : theta0 is the intercept when x=0



Our aim is to find the minimal cost line and to create a cost function The best red line possible

The distance in the submission should be minimal see the green lines

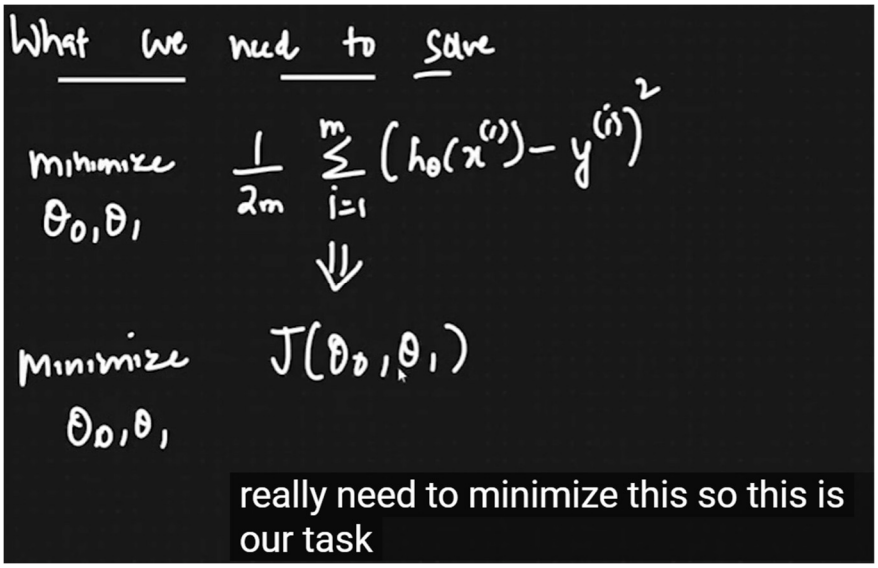



The entire equation is called as Squared error function, squaring is done to discard negative values

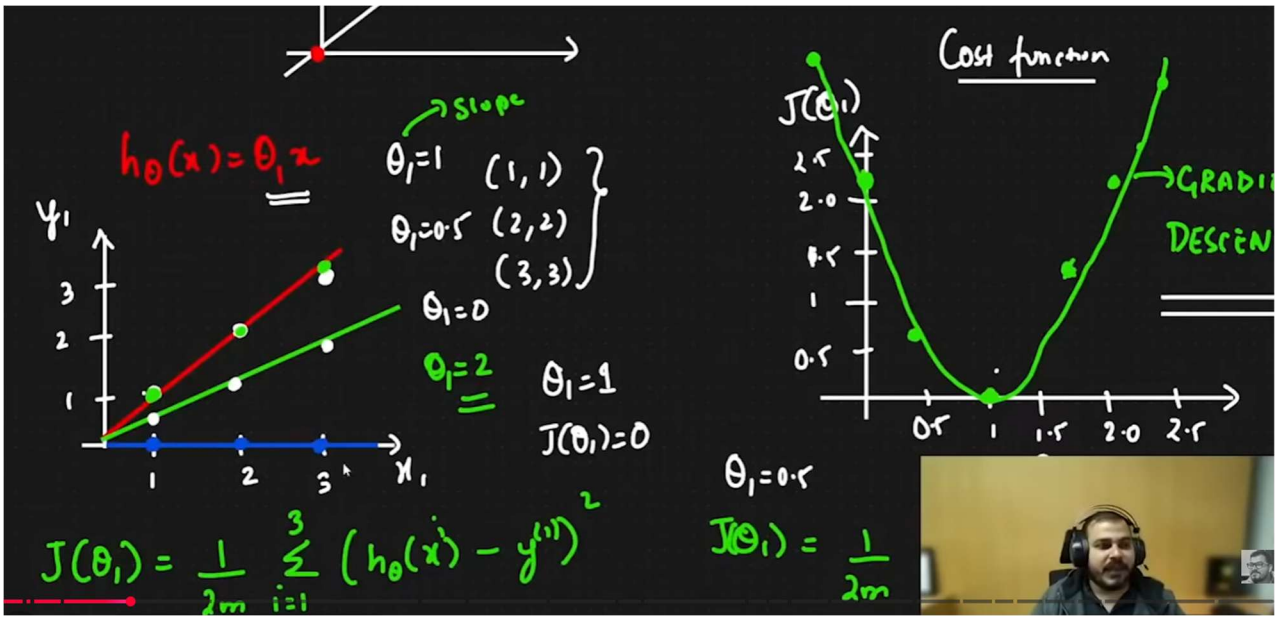
And m is the total number of points being compared, it is divided to get an average. It is divided by 2 for simplicity of mathematical equation

Time:30:00

What we need to actually solve

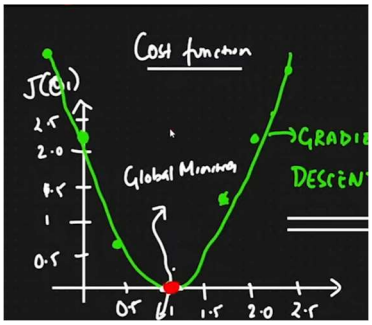



We need to make sure that we get the right theta 1 value

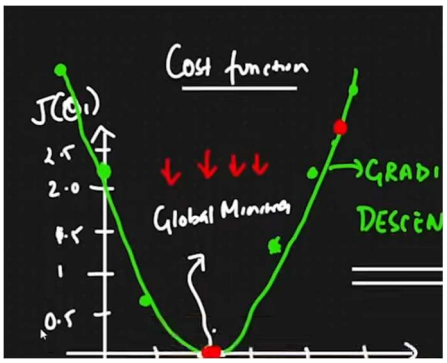



.

The best fit line represents global minimum

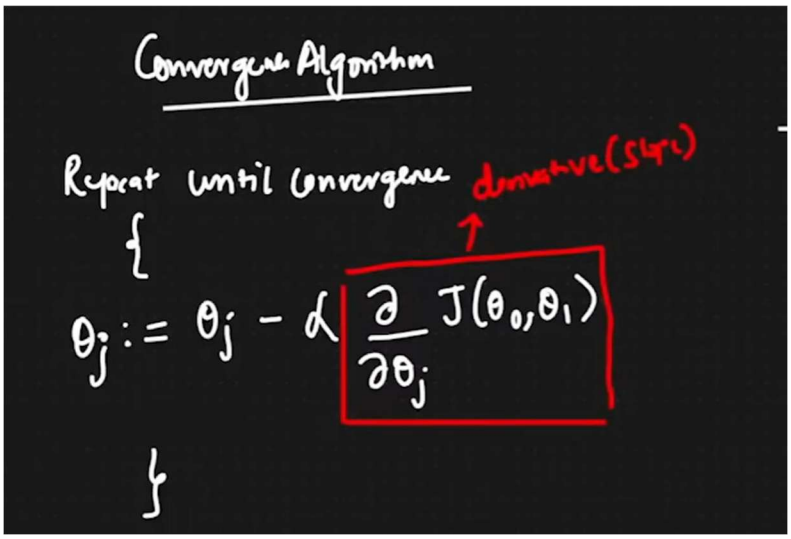



We need to try to find the red dot in the parabola that is in the upper section which can help us get to the global minimum



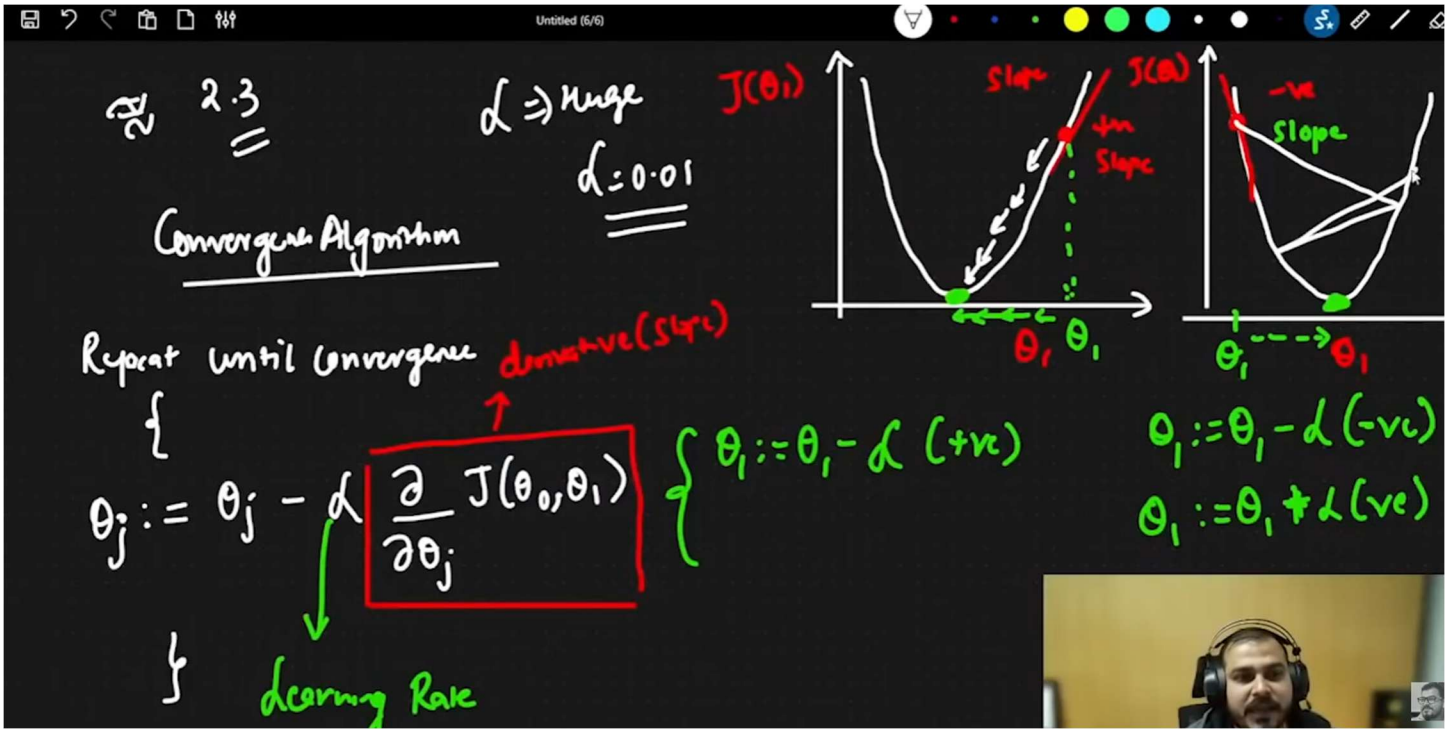

The Convergence Algorithm needs to be used to get to the

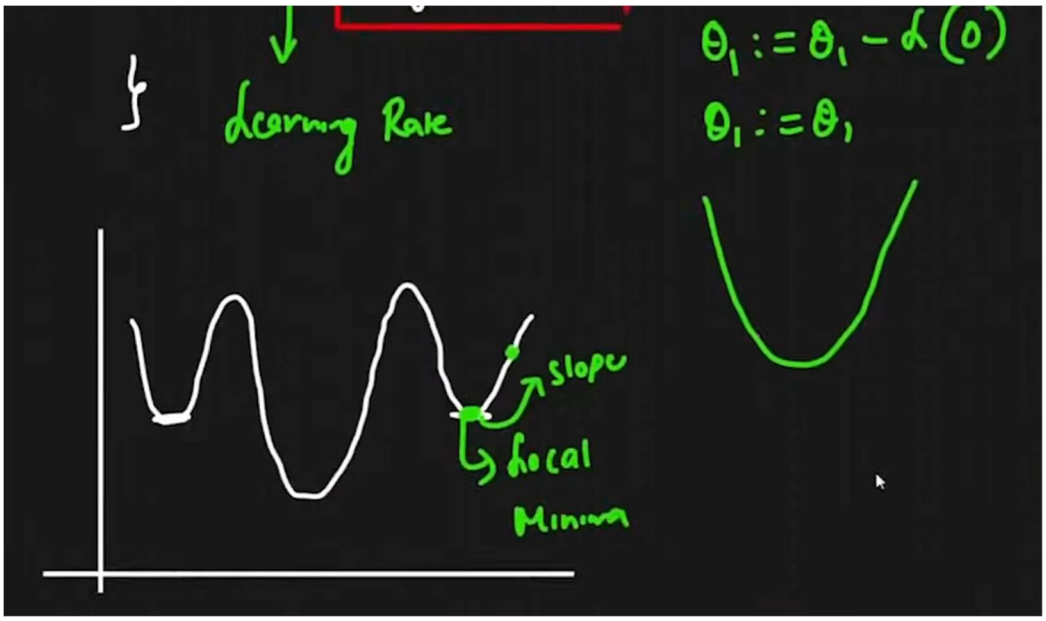
global minimum




How does it find itself ?
if we find the point in right hand side of the parabola or the right side of the parabola the formula will adjust itself relatively and also if the point is on the left side it will adjust accordingly .

Also we need to choose our alpha very carefully so that the points never reach the global minimum

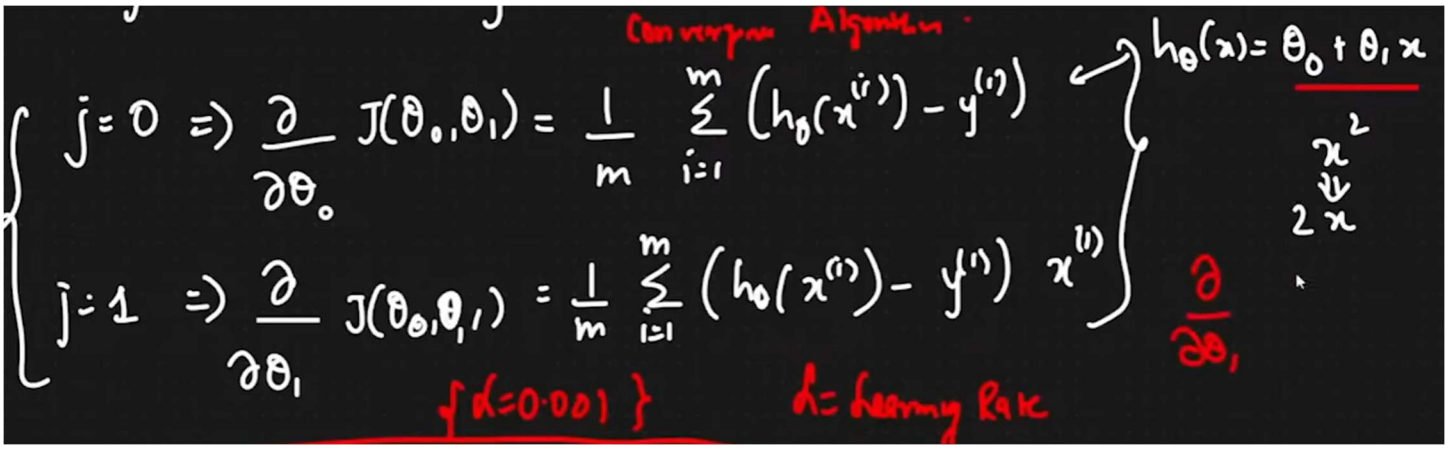


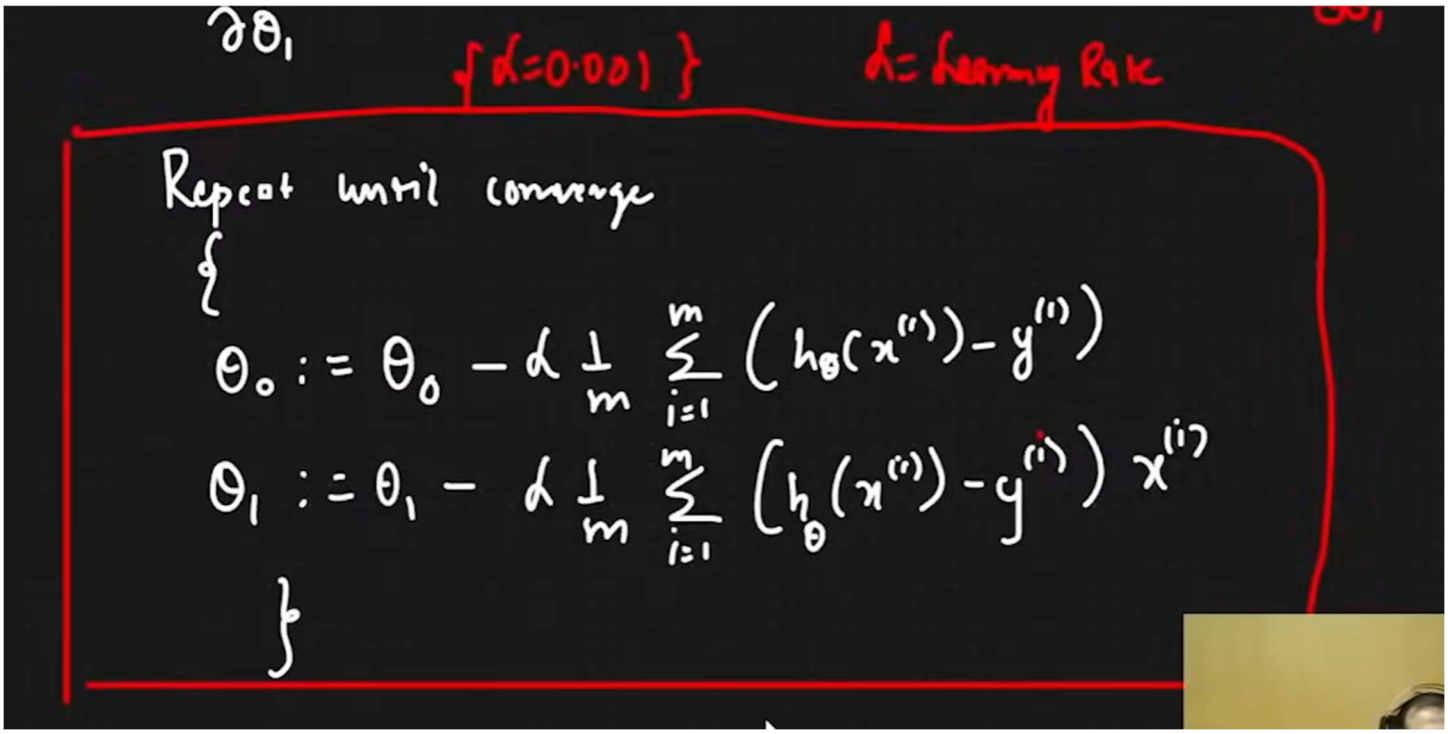




In some cases we can get the local minimum as the actual point, but this problem is not solved in linear regression and this could be solved using deep learning algorithms, Gradiennt descent and ANN in dl has lot of local dinima adam optimizers are used in such cases.

Gradient decent algorithm

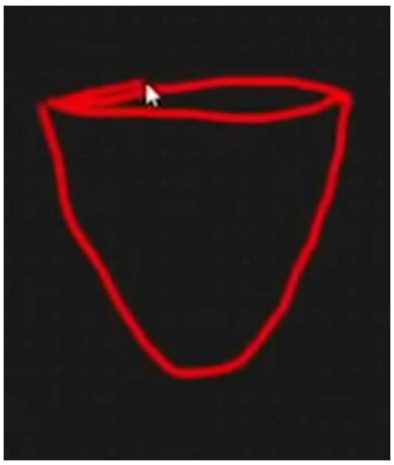
Repeat until convergence




55:08

We will have lot of of convex feature, at one point we will have a 3d curve: it will be coming down slowly just like you are descending from a mountain



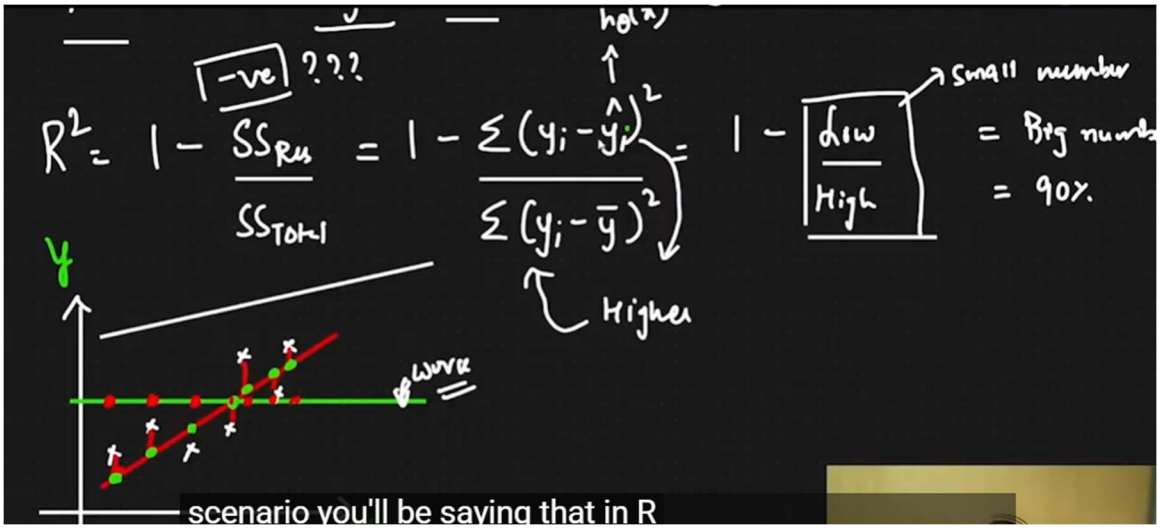
Two performance matrix

R62 and adjusted R^2

R^2 = 1- Sres/Stotal

Yihat : green points  YI IS THE ACTUAL POINT

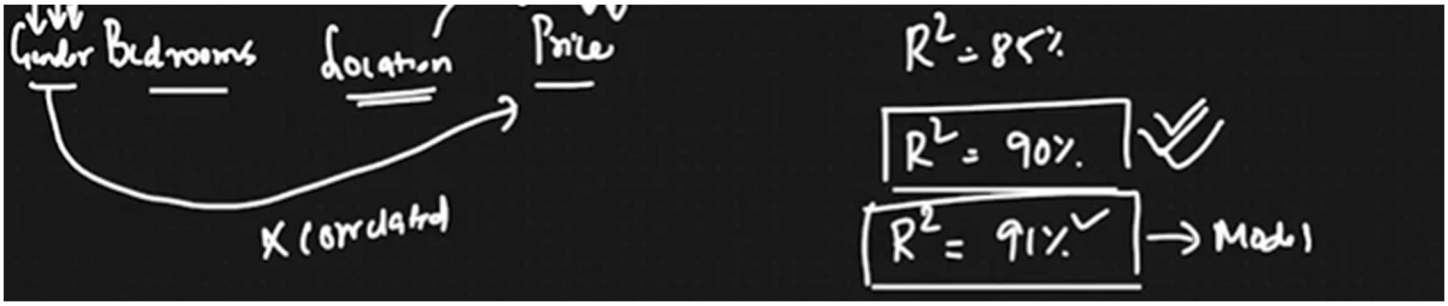
Y bar is the mean of y




If we keep on adding feature, then also there is a chance of increasing R^2, so in order to not increase

R^2 with irrelevant or related, so we need to fix that problem like the example here gender is not related at all with the pricing of the housing to increase or decrease, so we need to fix that issue with this corrected

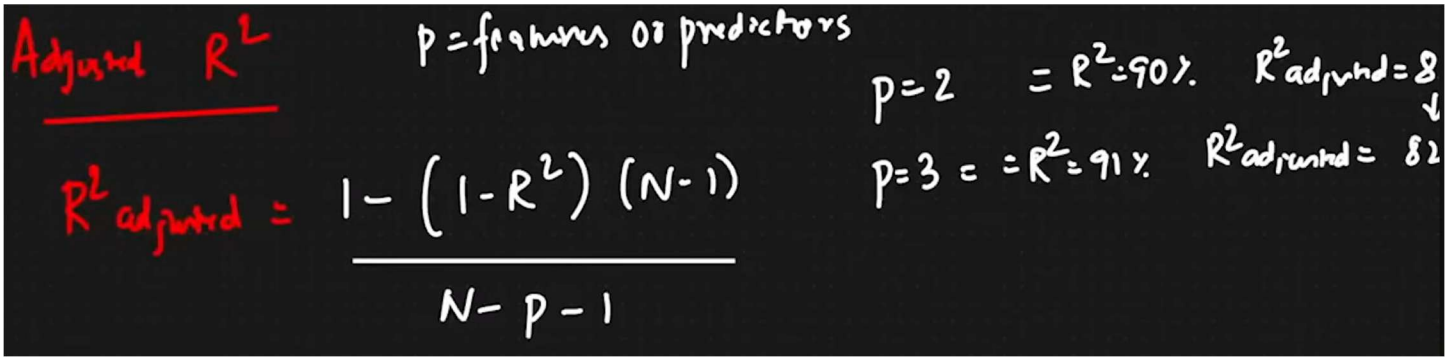
formula




Adjusted R^2 = 1- ((1-R62)(N-1))/(N-P-1)

Note: N is the total number of samples, P is the number of features/predictors

When the features are corelated, the r^2 value will be greater than that of non corelated values



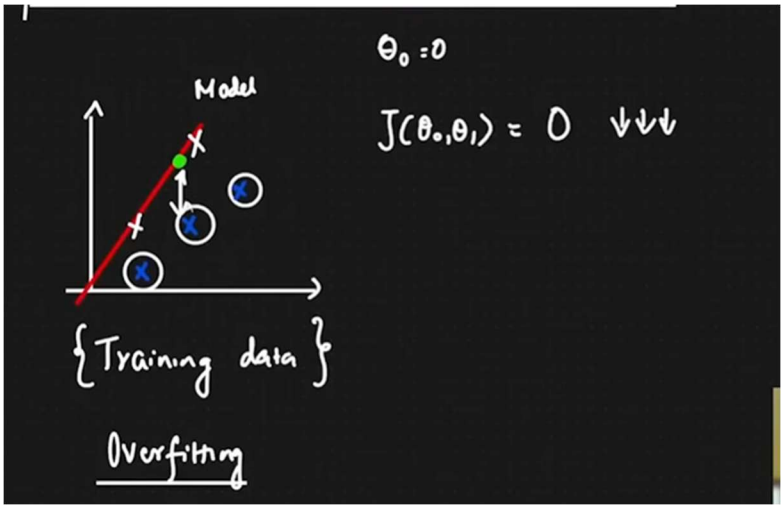

Interview Question: What is bigger R^2 or adjusted R^2

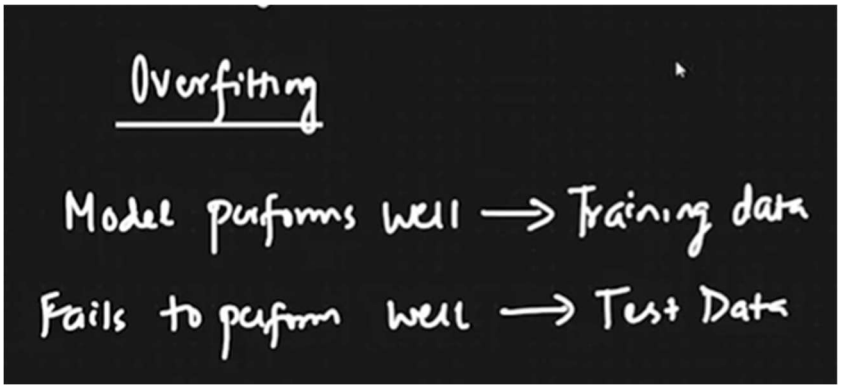
Ans: We know that R^2 will be bigger because R^2 does not take consideration of the correlation of the feature/Predictor

- **Ridge And Lasso  Regression**
- Assumption of Linear Regression
- Logistic Regression
- Confusion Matrix
- Practicals for Linear Ridge Lasso and Logistic Regression

# Ridge And Lasso  Regression

Even though my model has trained well with the training data, it causes overfitting, which means my model performs well with training data and fails to perform well with the test data



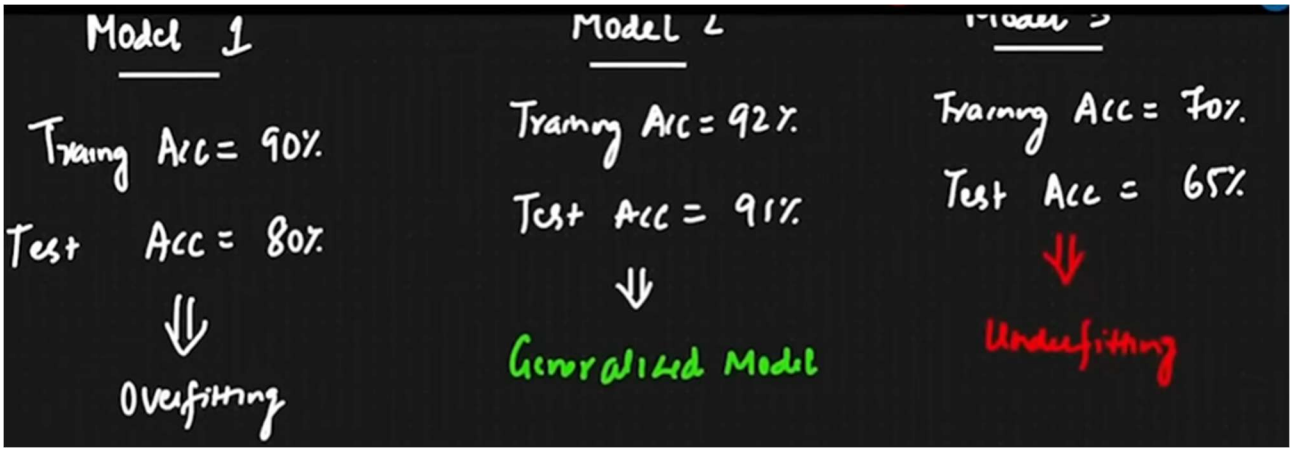

**Overfitting**

- When the model performs well with training data: Low Bias
- When it fails with the test data: High variance

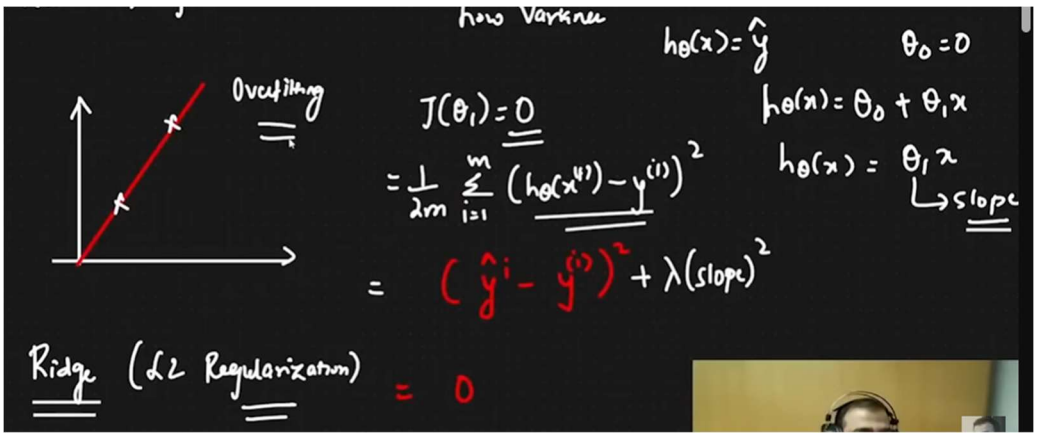
**Underfitting**

- Model accuracy is bad with training data: ( High Bias )
- Model accuracy is also bad with testing data: (High Variance)

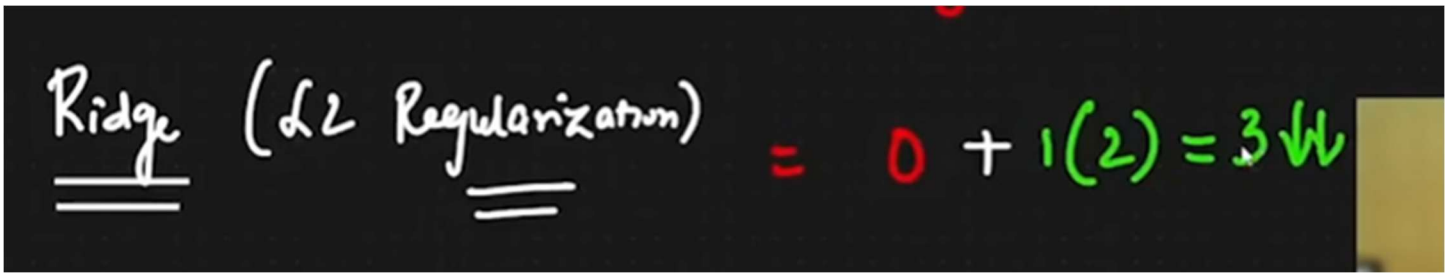
So here is an example of different senarios :




IN our example we used the equation and we trued to predict the J with it, and it has caused an overfitting condition, so to change that 0 into something else we need ti use Ridge regression also known as the L2 regularization

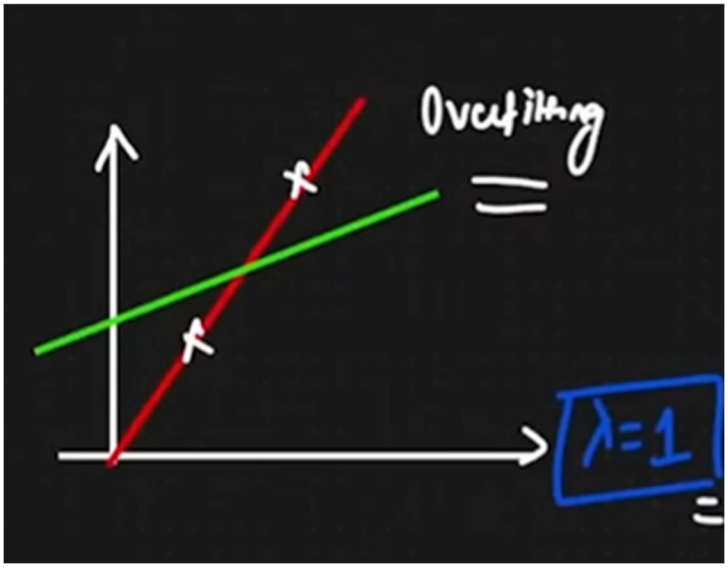



Let us consider our lamda as 1 and consider slope as 2 and the result becomes 3(Lamda is the hyper parameter)



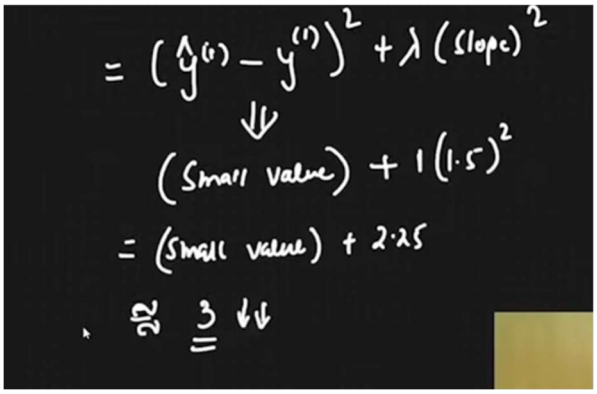

So due to this it will try to again change the best fit line to something else, and therefore it will change the theta 1 value

Now to change that we will again change the best fit line equation to something else :




Let us consider the value of slope to be changed from 2 to 1.5

This is why we are adding ridge L2 regularization, as the value is decreasing:



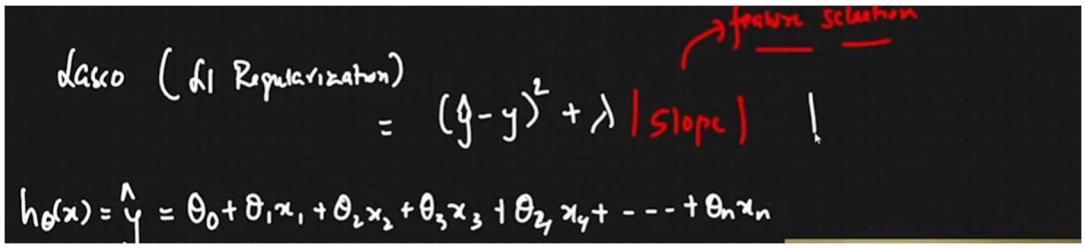
- This is used to prevent overfitting
- The steep should not be steep, and it should hep us create a  generalized value
- Now we might have to specify the iterations{Hyperparameter}
- We can not get 0 because it could be an overfitting model

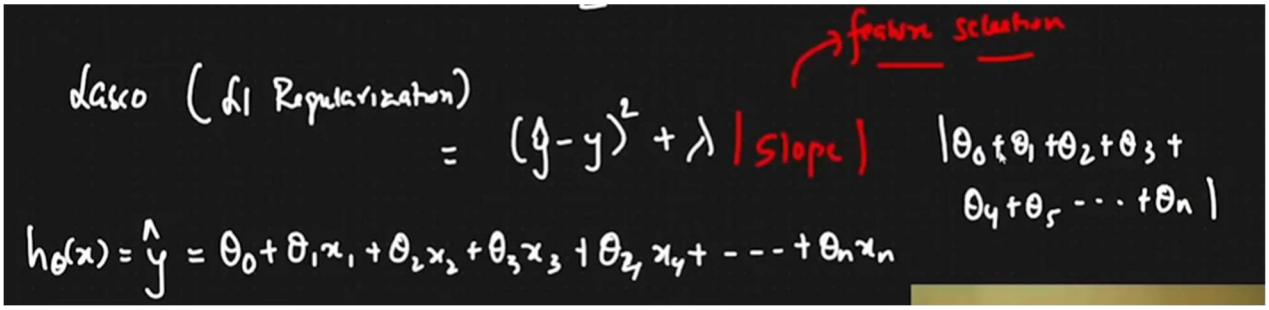
Lamda: How fast we can grow or lessen the steepness, this is changed by checking the hyperparameter, the examples will be shown later in the notes.

---

Lasso Regularizartion = (yhat-y)^2  + lamda |slope|

- Mod of slope will help us do feature selection



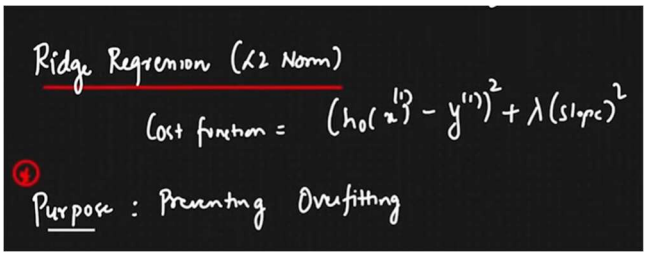

We will try to remove those feature which are not at all useful in this problem statement


Lamda basically means cross validation, cross validation is the technique where we will try to train our model

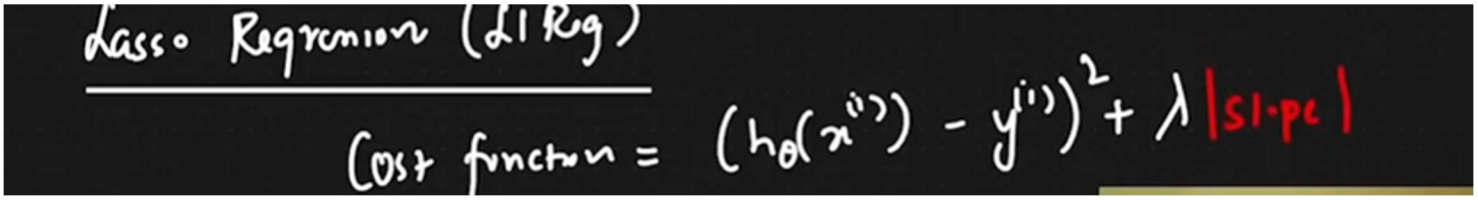
- **In short we are trying to reduce the cost function on the basis of lamda and slope value**

  **We should try both algorithms and see which has the best regressuin funcgtion with different cost function**

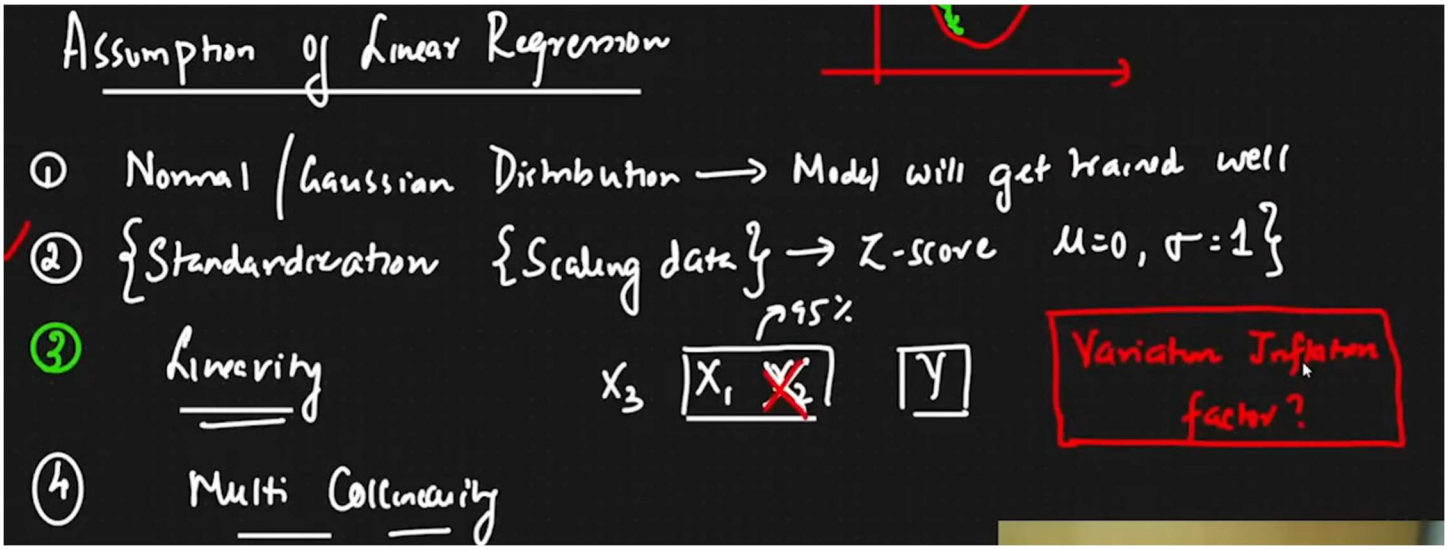
- **Ridge regression prevents overfitting**




- **Lasso regression prevents users from using unnecessary features/ help in feature selection**

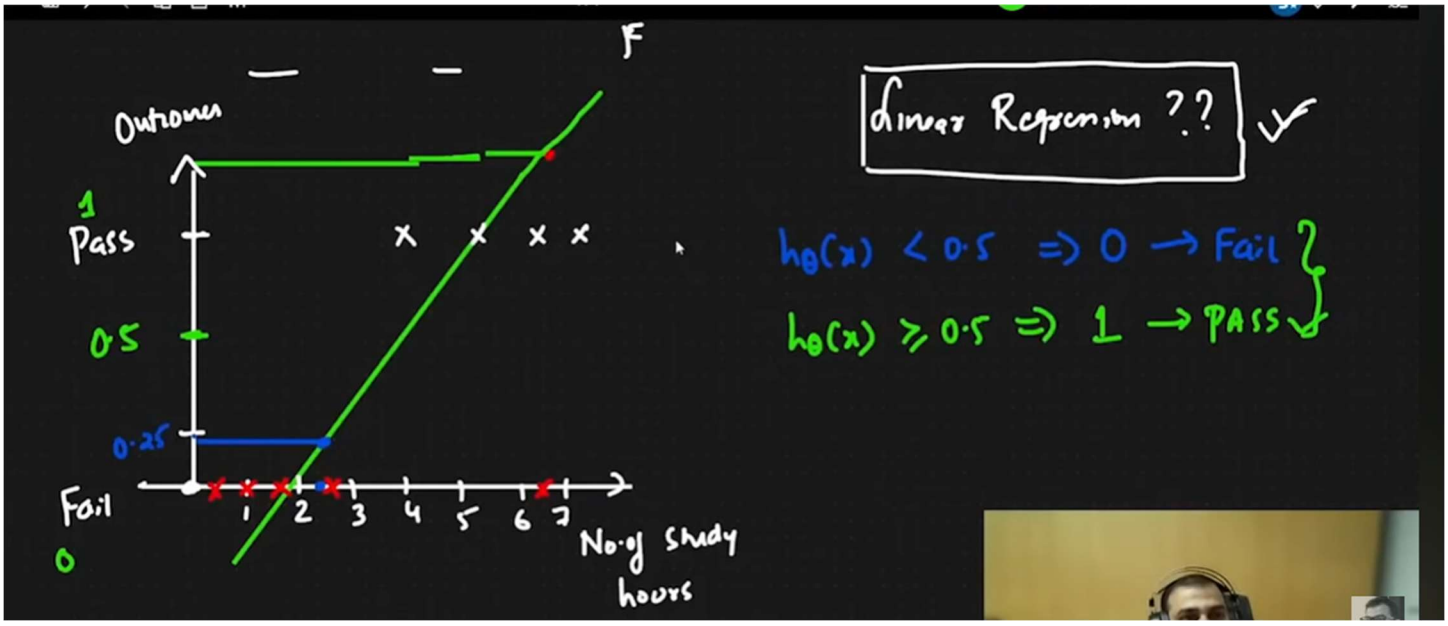



---



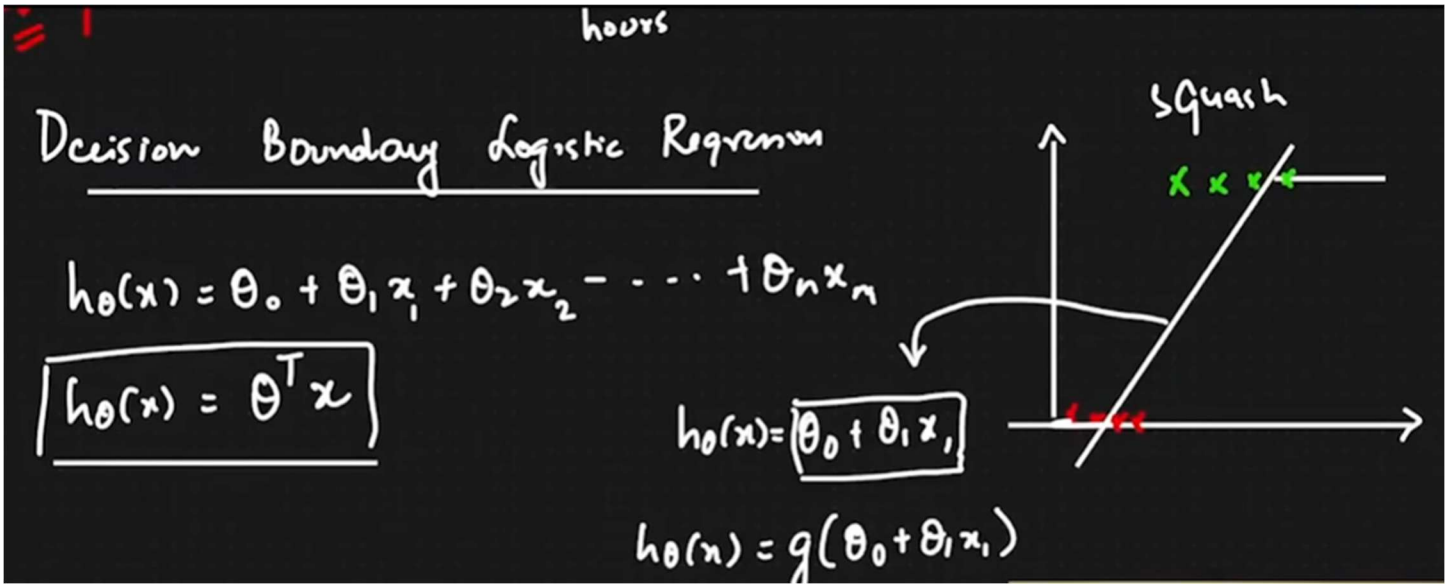

---

LOGISTIC REGRESSION is the first example for classification -> works very well with Binary Classification




To pass >3(basic logic)

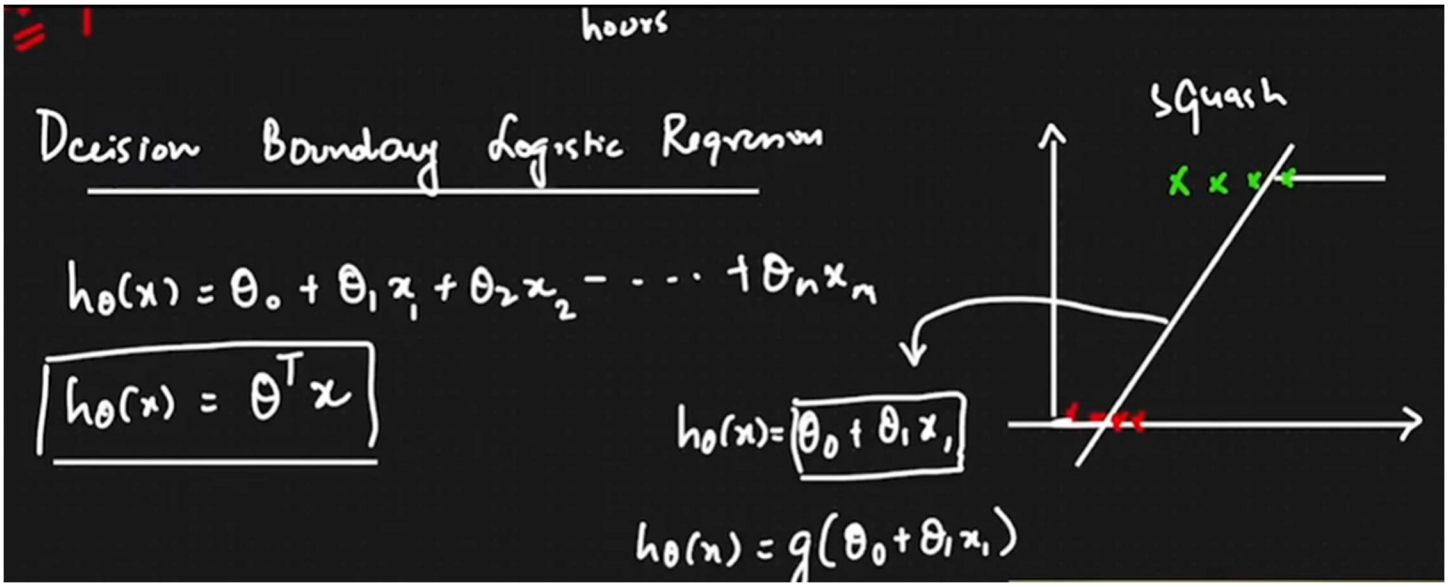
- Why not use linear regression then ?
  - With outlier the entire bestfit line would change, now for even 5 hrs study some people could fail according to this new model
- We could have to squash the functionality and this would be not good. Sepcially for negative inputs



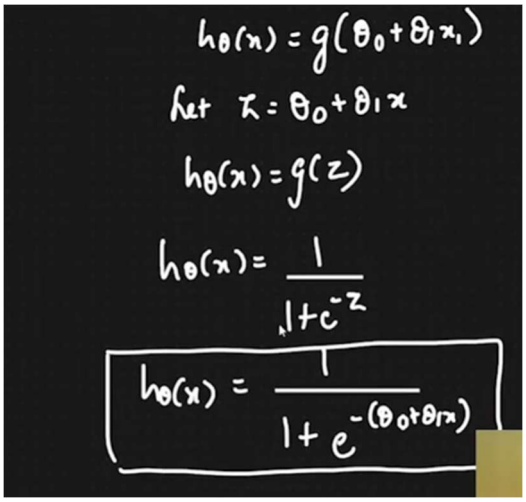
BINARY CLASSIFICATION

- Our output value should always be 0 or 1

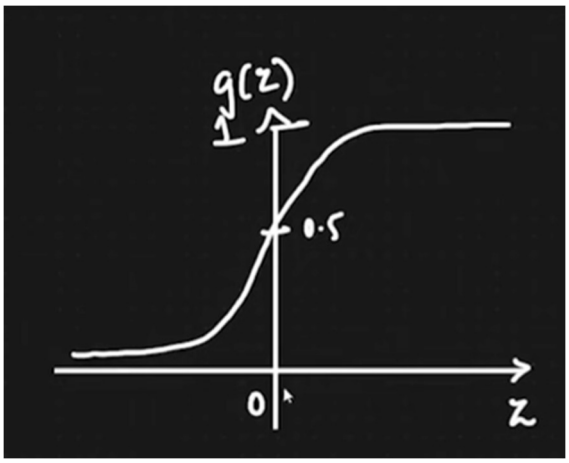
Decision Boundary Logistic regression



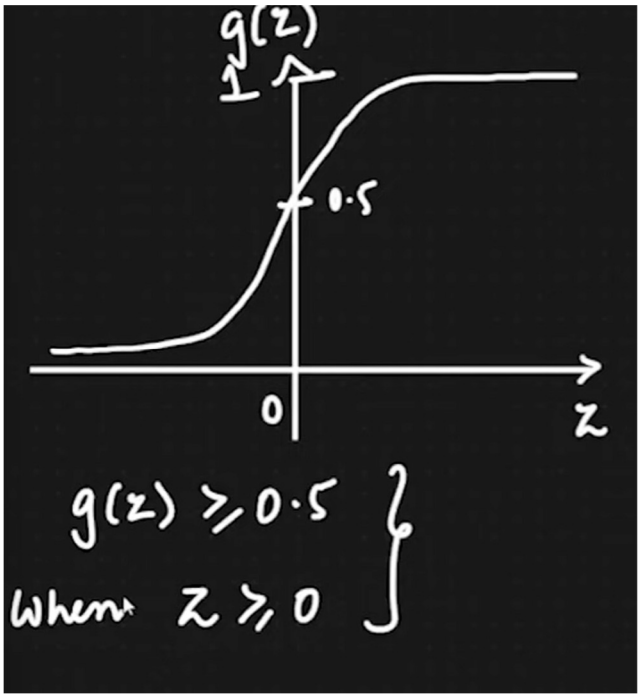

We will try to apply a mathamtical formula which would squash this line, it is also called sigmoid/Logistic activation function

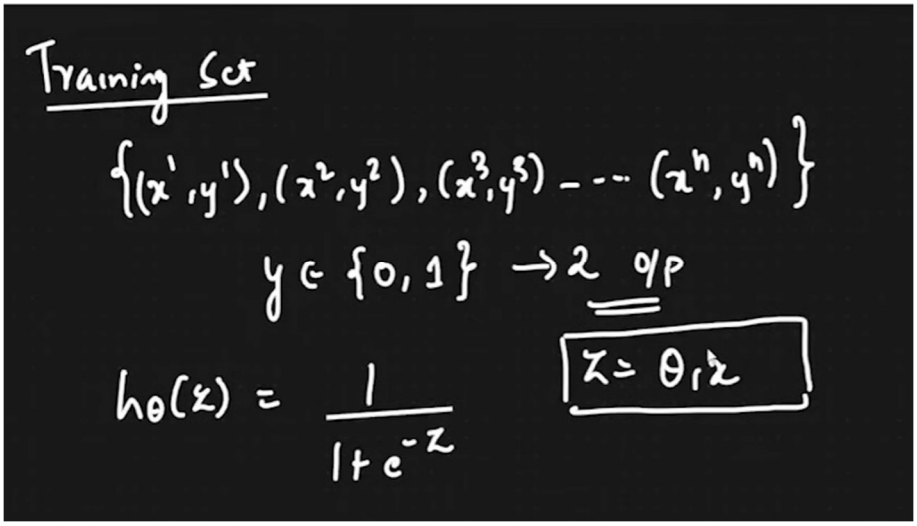


Now this will be able to squash the regular line to desired value / this function is also known as logistic or sigmoid function and this will look something like



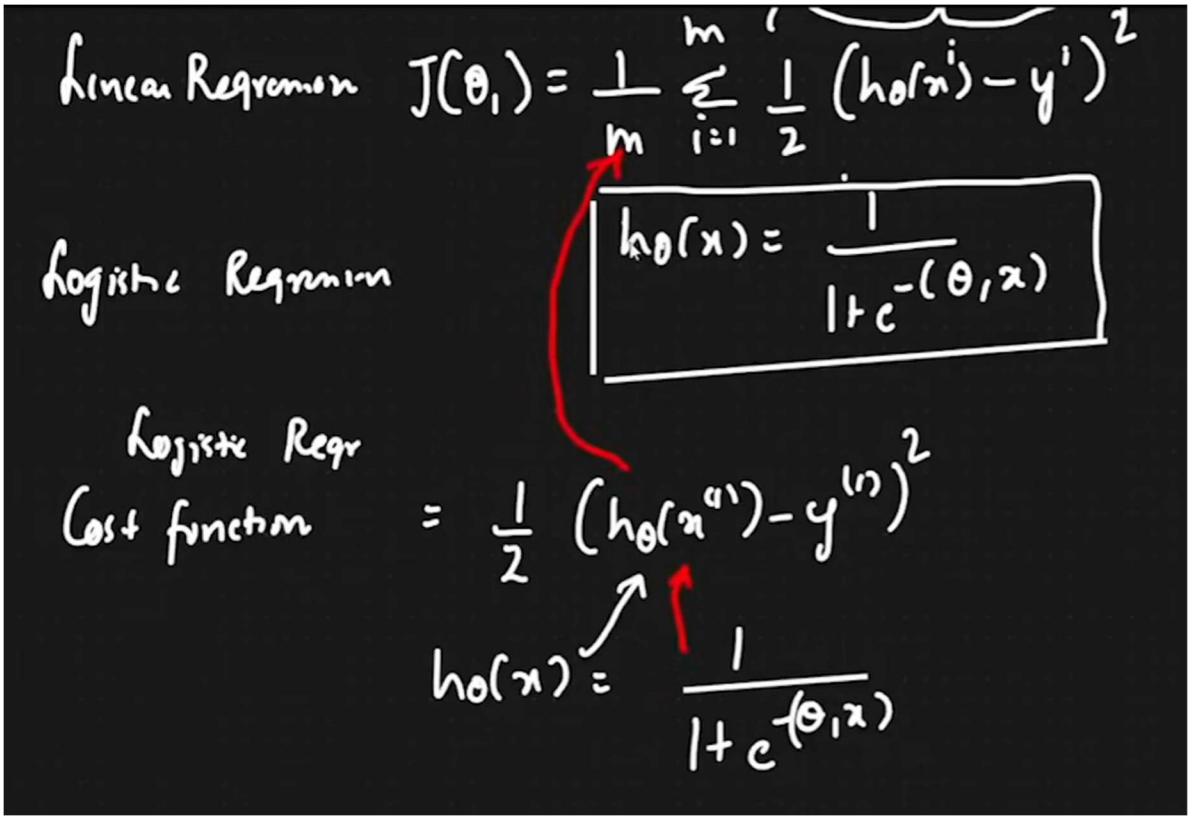

So we can make the assumption that





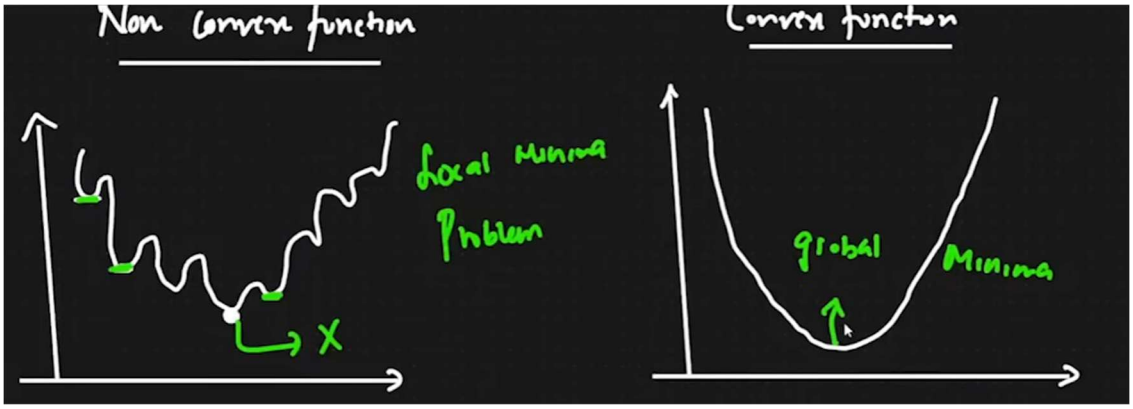

Change parameter theta 1 to get the best fit line and the sigmoid activation formula

Cost function =



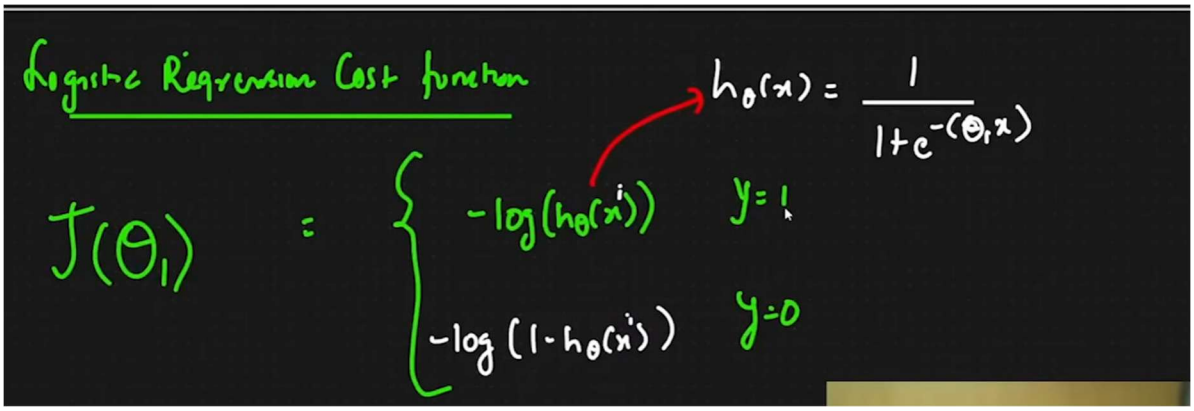

This can not be used because it is a non convex function:

| Non Convex Function | Convex Function |
|---|---|
| Lots of local minima | No local minimas |
| Never reach global minima | Reach global Minima |

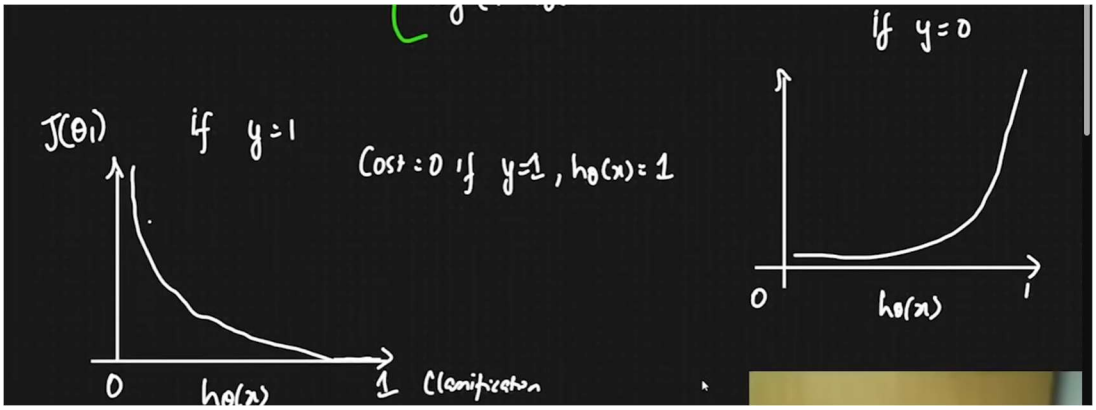



Logistic regression cost function:

New type of CF is suggested=



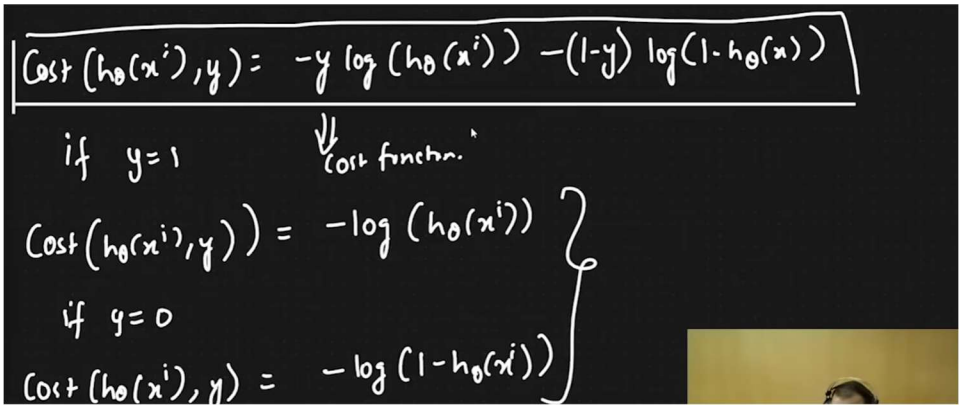

So if y-1, h(theta) =1 then the the left side graoh is made and the right side graph is for when y=0

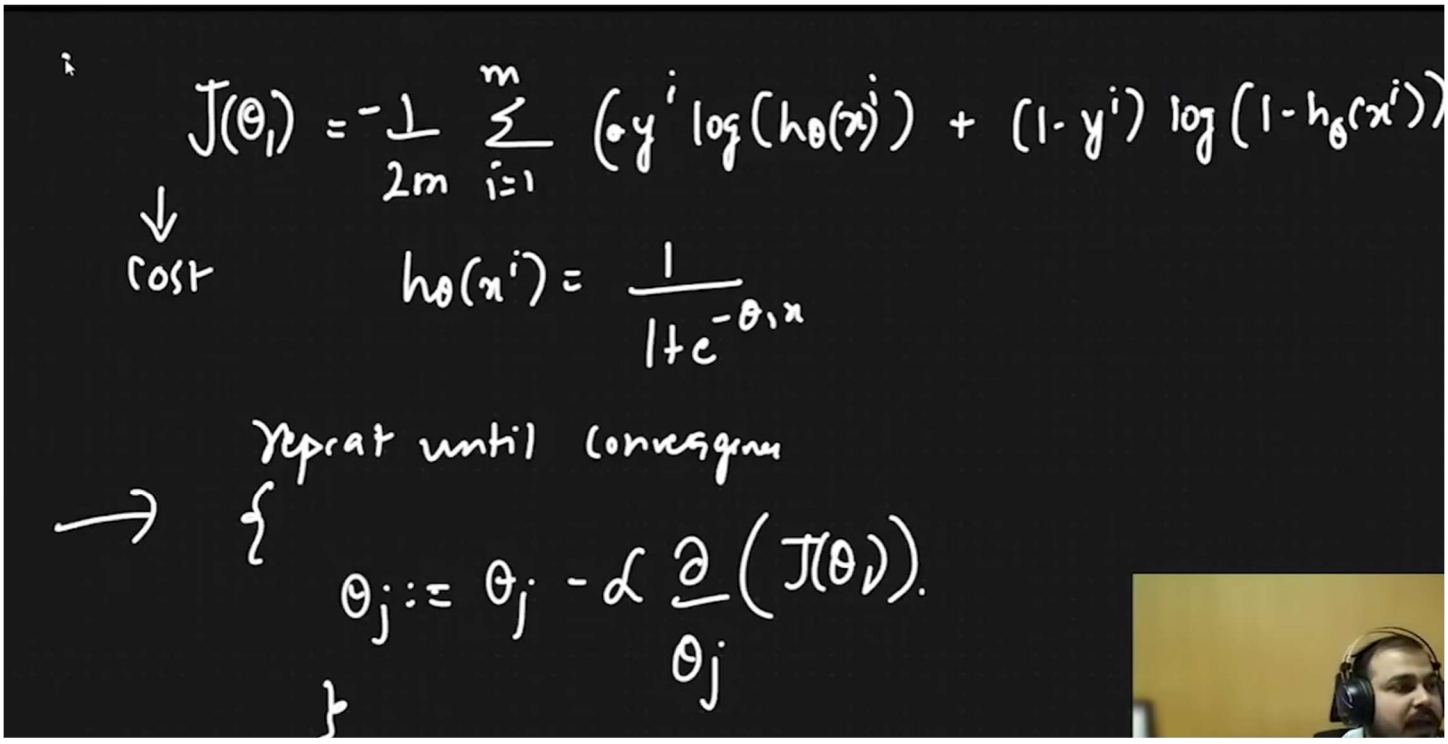


Now if we combine both we will get a convex function and the final cost function will be :

$$\text{Cost}(h_\theta(x^i), y) = -y \log(h_\theta(x^i)) - (1-y) \log(1 - h_\theta(x))$$

Ig we replace y with 1 then you will only get a particylar value :
If y will be 0 then the following function will be executed:

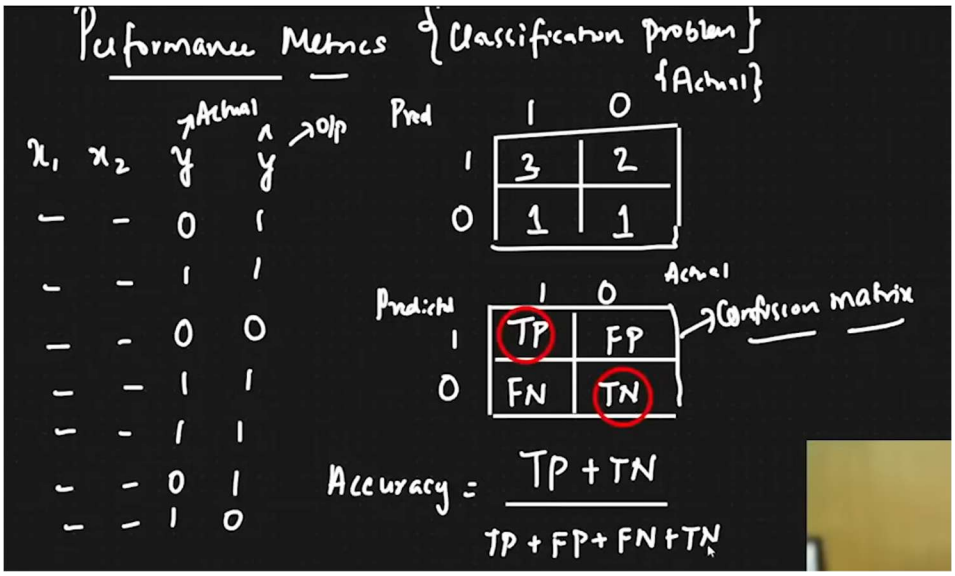


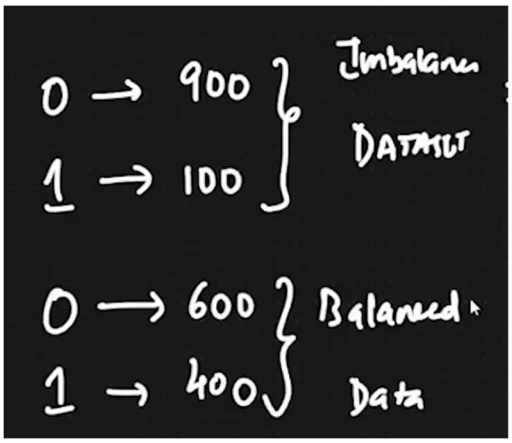



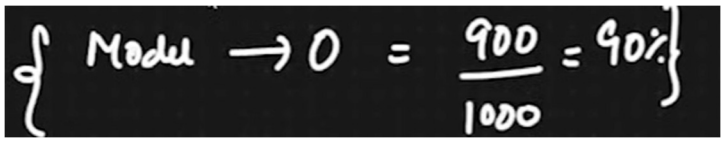
Perfprmance Measure

TP represents true positive
TN represents true negative

FP represents False Positive
FN represents False Negative



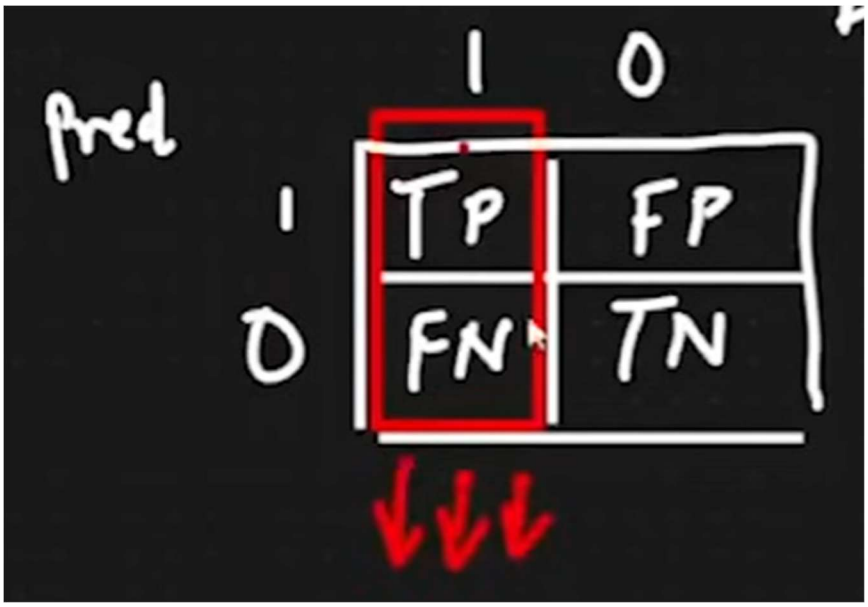




If there is imbalanced data set then there will be problems in the accuracy and the algorithm, so we can use

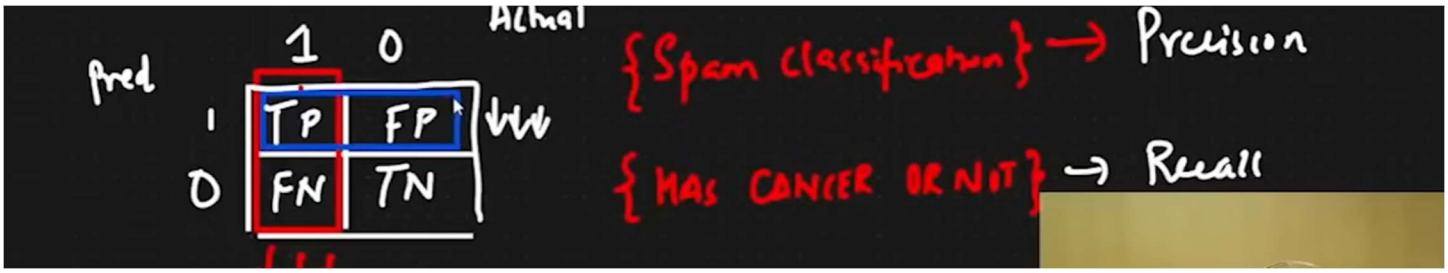
- **Precision**
  - TP/(TP +FP)

- **Recall**: Out of all true positives how many of them have predicted positive, now we try to reduce False negatives

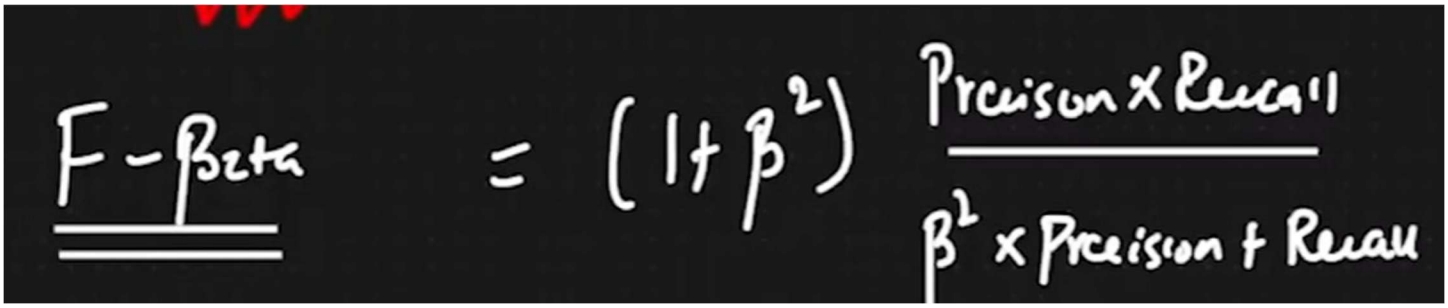



RECALL = TP/(TP+FN)

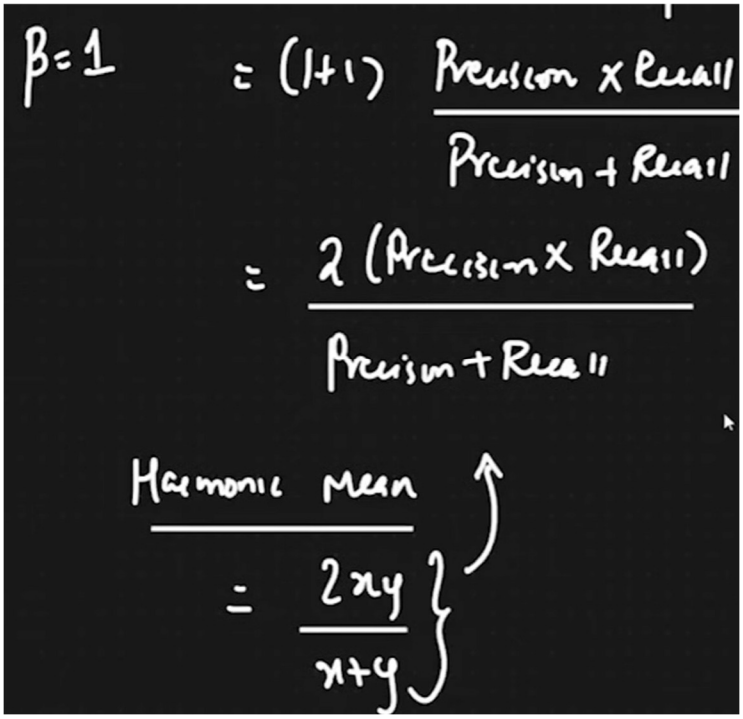
Precision means False positive is given more priority to be reduced



**F-Beta/F-score(3 different formula) The generic one is also termed as Harmonic mean**

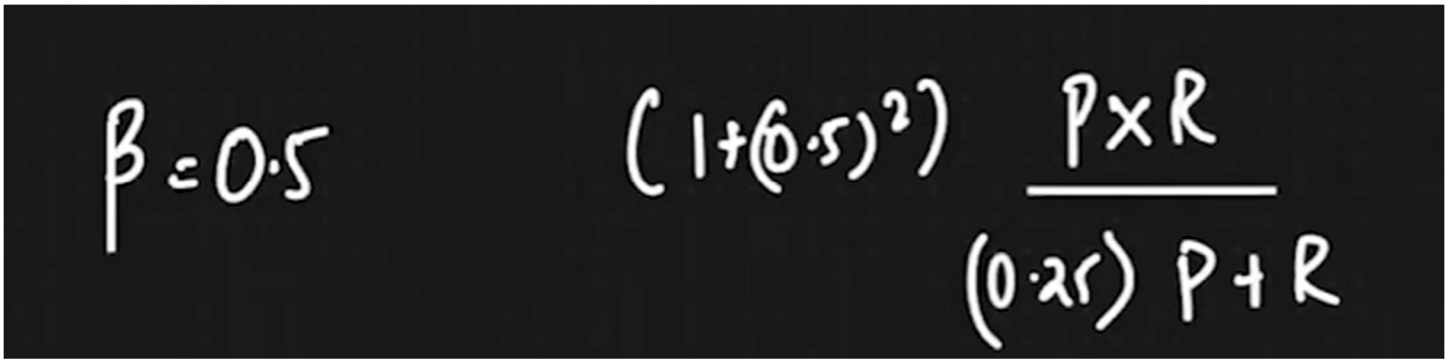



For the generic case we have the following explanation :




Lets say that the false positive is more important than false negative, so we consider the B as 0.5, so decreasing means we are creating more importance on False positive than false negative:

**F0.5 score is used**



Now if we do the other thing it becomes the **F2 score**


DAY3: Please look at the .pynb file in github from time 2:35 continue with Lasso regression to see the results

---

Agenda:

# Naive Bayes Algorithm (Classification)

↳ Probability and Bayes Theorem

## Independent Event

Rolling a Dice $\{1, 2, 3, 4, 5, 6\}$

$P(1) = \frac{1}{6}$ $P(2) = \frac{1}{6}$ $P(3) = \frac{1}{6}$

## Dependent Event

What is the probability of removing a white marble and then a yellow marbles?


↳ $P(w) = \frac{3}{5}$ → 1st Event


$P(y/w) = \frac{2}{4} = \frac{1}{2}$ → 2nd Event

↳ Conditional probability

$P(w \text{ and } y) = P(w) * P(y/w)$ ⇒ Independent Event

↑ ↑ ↳ Conditional probability

# Bayes's Theorem

$$P(A \text{ and } B) = P(B \text{ and } A)$$

$$P(A) * P(B/A) = P(B) * P(B/A)$$

$$\boxed{P(B/A) = \frac{P(B) * P(B/A)}{P(A)}} \Rightarrow \text{Bayes's Theorem}$$

Dataset — Independent feature :- → O/P

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | ---- | $x_n$ | $y$ |
|---|---|---|---|---|---|---|

$$P(y/x_1, x_2, x_3 \cdots x_n) = \frac{P(y) * P(x_1, x_2, x_3, x_4 \cdots x_n/y)}{P(x_1, x_2, x_3 \cdots x_n)}$$

$$= \frac{P(y) * P(x_1/y) * P(x_2/y) * P(x_3/y) \cdots * P(x_n/y)}{P(x_1, x_2, x_3, \cdots x_n)}$$

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|
| — | — | — | — | Yes |
| — | — | — | — | No |
| — | — | — | — | Yes |

$$P(Yes/x_1, x_2, x_3, x_4) = \frac{P(Yes) * P(x_1/Yes) * P(x_2/Yes) * P(x_3/Yes) * P(x_4/Yes)}{\text{Constant} \leftarrow P(x_1) * P(x_2) * P(x_3) * P(x_4)}$$

$$P(No/x_1, x_2, x_3, x_4) = \frac{P(No) * P(x_1/No) * P(x_2/No) * P(x_3/No) * P(x_4/No)}{\text{Constant} \leftarrow P(x_1) * P(x_2) * P(x_3) * P(x_4)}$$

## Lets Solve this Problem

| Day | Outlook | Temperature | Humidity | Wind | Play Tennis |
|---|---|---|---|---|---|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rain | Mild | High | Weak | Yes |
| 5 | Rain | Cool | Normal | Weak | Yes |
| 6 | Rain | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Strong | Yes |
| 8 | Sunny | Mild | High | Weak | No |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rain | Mild | Normal | Weak | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rain | Mild | High | Strong | No |

Outlook

| | Yes | No | P(E/Yes) | P(E/No) |
|---|---|---|---|---|
| Sunny | 2 | 3 | 2/9 | 3/5 |
| Overcast | 4 | 0 | 4/9 | 0/5 |
| Rain | 3 | 2 | 3/9 | 2/5 |

Temperature → Test (Sunny, Hot) → O/P = PLAY (Y)

| | Yes | No | P(E/Yes) | P(E/No) |
|---|---|---|---|---|
| Hot | 2 | 2 | 2/9 | 2/5 |
| Mild | 4 | 2 | 4/9 | 2/5 |
| Cool | 3 | 1 | 3/9 | 1/5 |

Yes 9 → P(Yes) = 9/14

No 5 → P(No) = 5/14

$$P(Yes/(Sunny, Hot)) = \frac{P(Yes) * P(Sunny/Yes) * P(Hot/Yes)}{\text{Constant} \leftarrow P(Sunny) * P(Hot)}$$

$$= \frac{9}{14} * \frac{2}{9} * \frac{2}{9}$$

$$= \frac{2}{63} = 0.031$$

$$P(No/(Sunny, Hot)) = \frac{P(No) * P(Sunny/No) * P(Hot/No)}{X \leftarrow P(Sunny) * P(Hot)}$$

$$= \frac{5}{14} * \frac{3}{5} * \frac{2}{5}$$

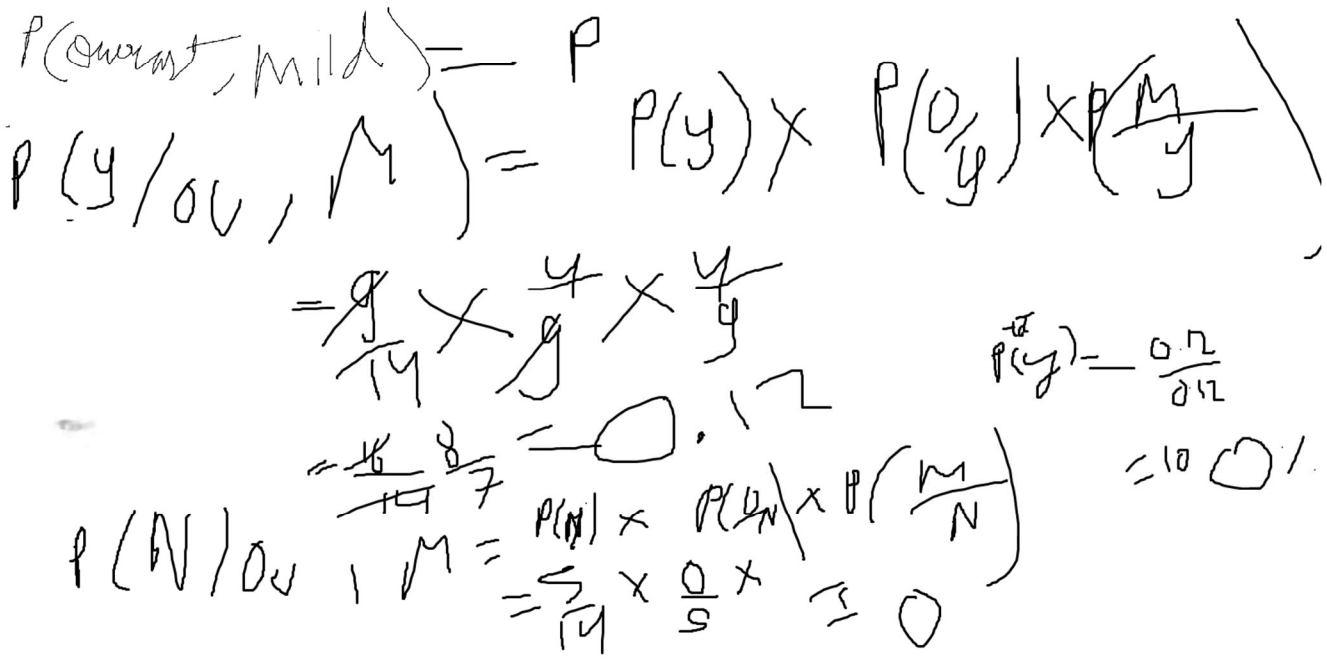
$$= \frac{3}{35} = 0.085$$

$$P(\text{Yes}/(\text{Sunny, hot})) = \frac{0.031}{0.031+0.085} = 0.27 = 27\%$$

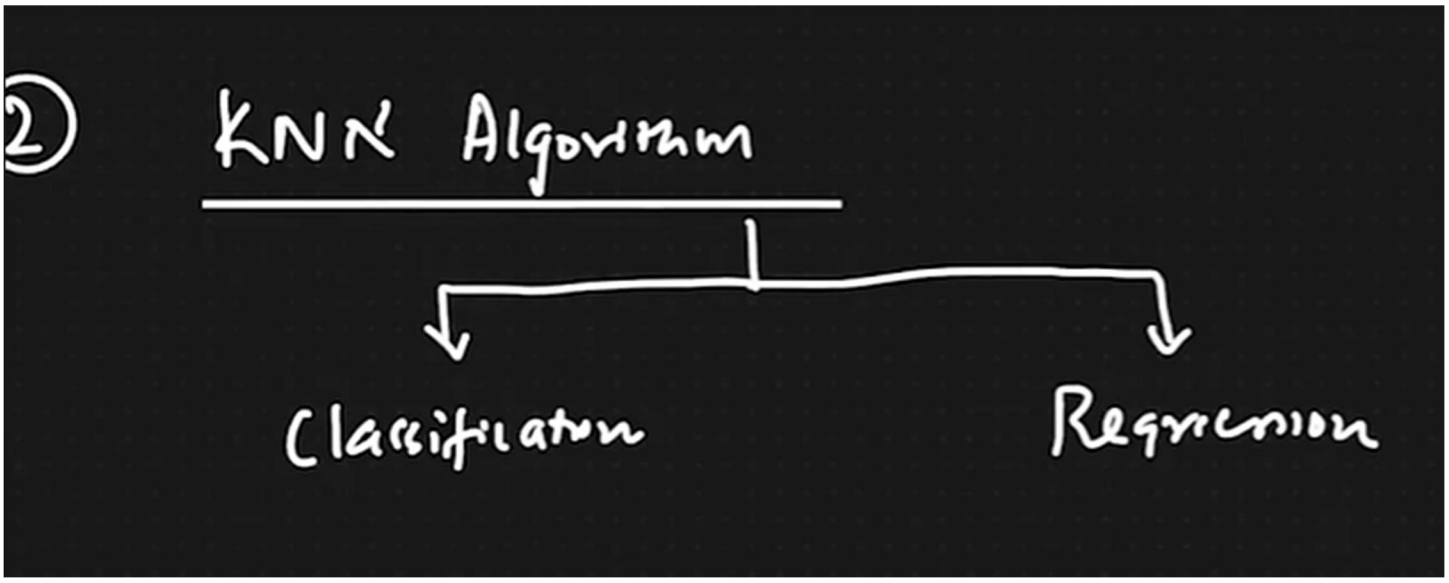
$$P(\text{No}/(\text{Sunny, hot})) = \frac{0.085}{0.031+0.085} = 0.73 = 73\%$$

Test data (Sunny, Hot) = 73% They will not play Tennis

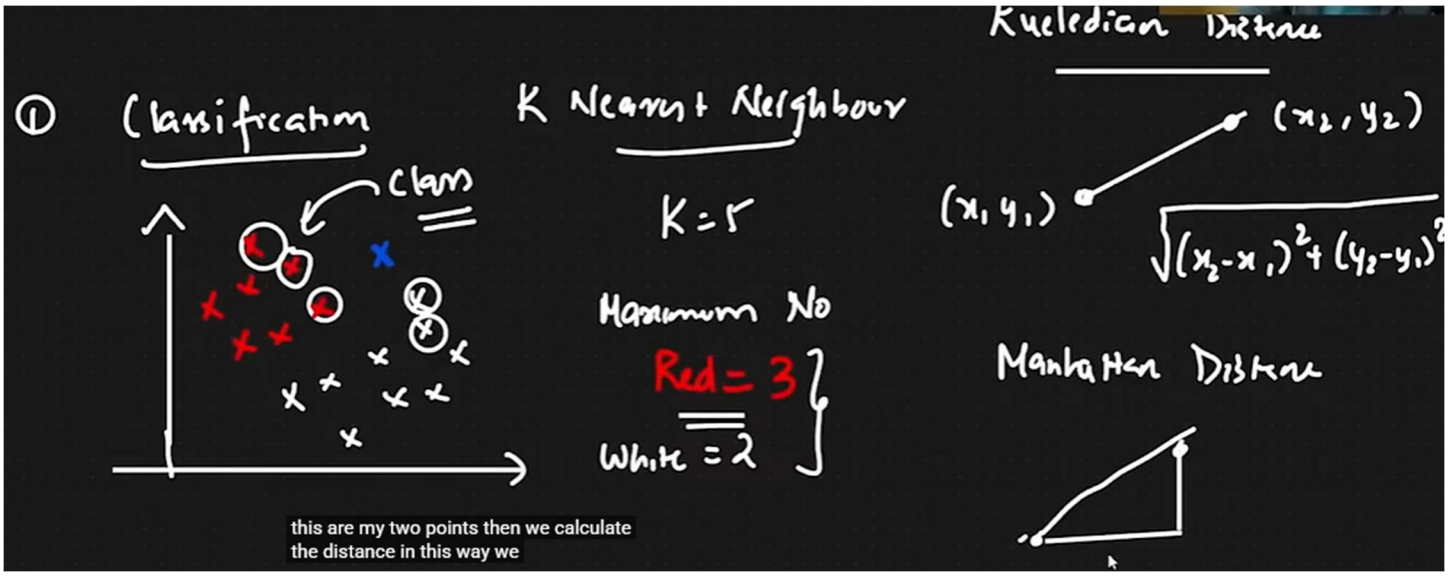
⇒ Person is not going to play Tennis



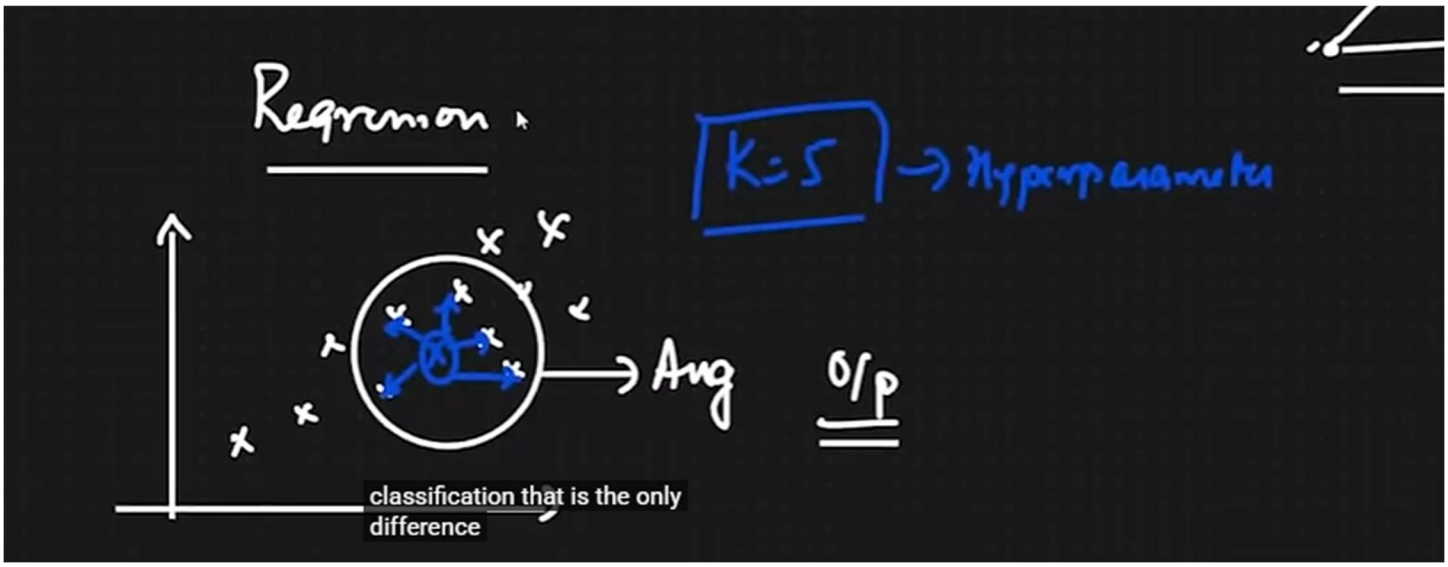
KNN -> K nearest neighbor



Suppose a new data For Classification (binary)

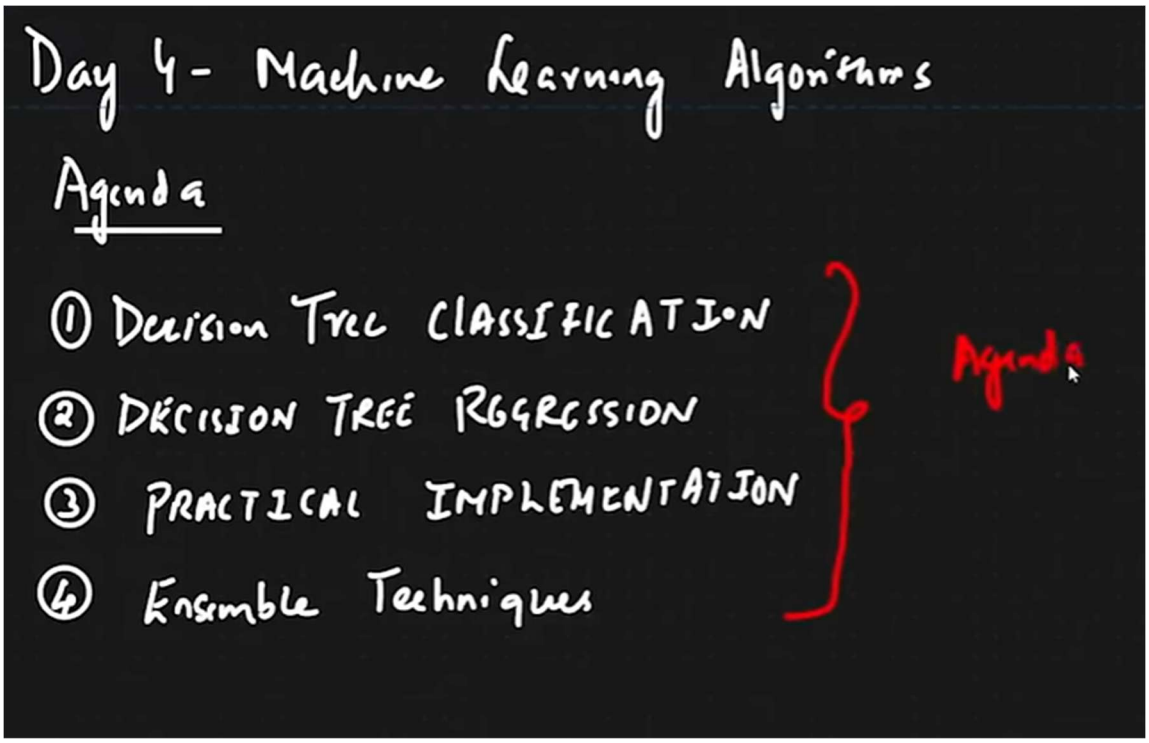


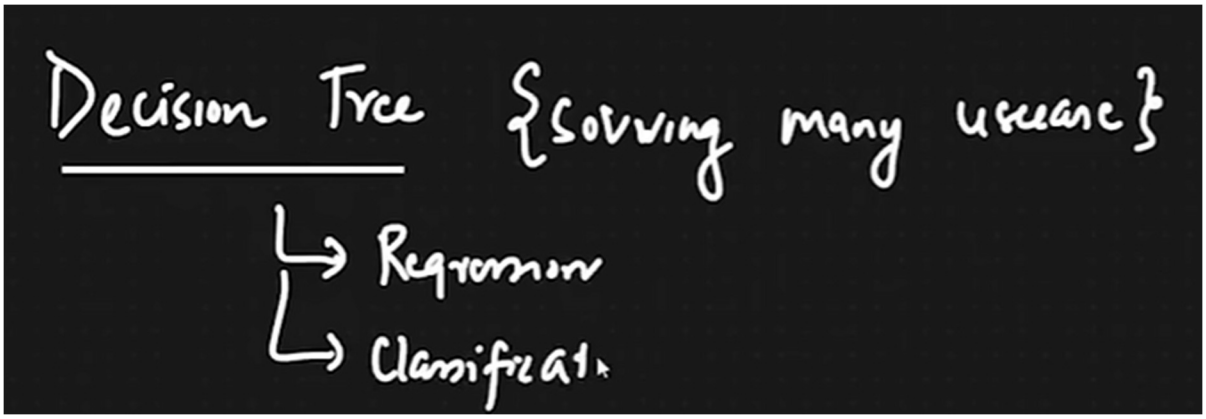

For Regression what do we do ?




Try K 1 to 50, and if the error rate is low that is how we choose the output

K nearest neighbor works very bad with outlier and with Imblalanced Data set example




If(age<18):

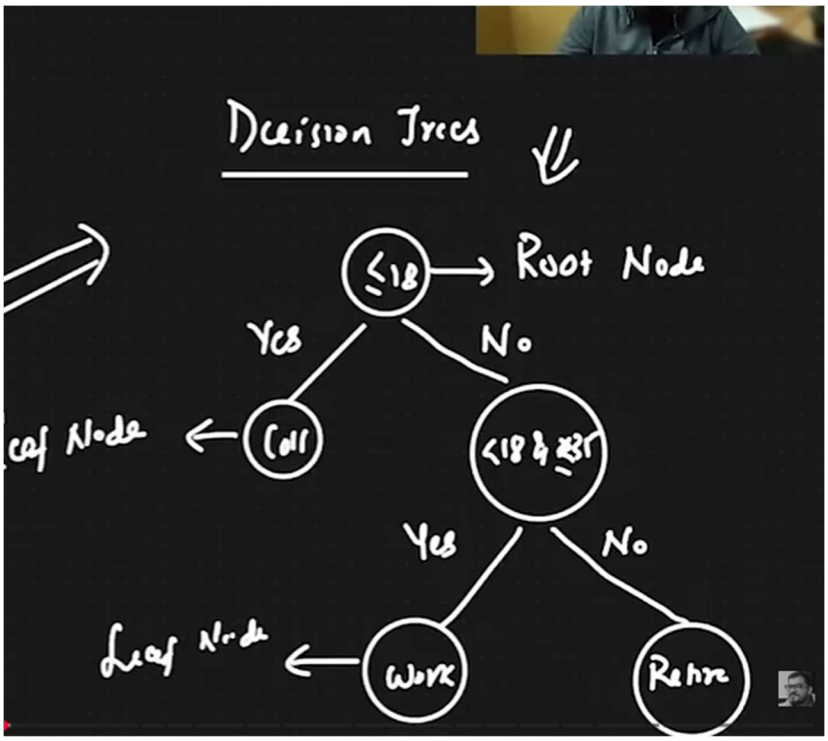
Print(college)

Elseif (age>18 and age<= 35):

Print (“worl”)

Else print(“retire”)

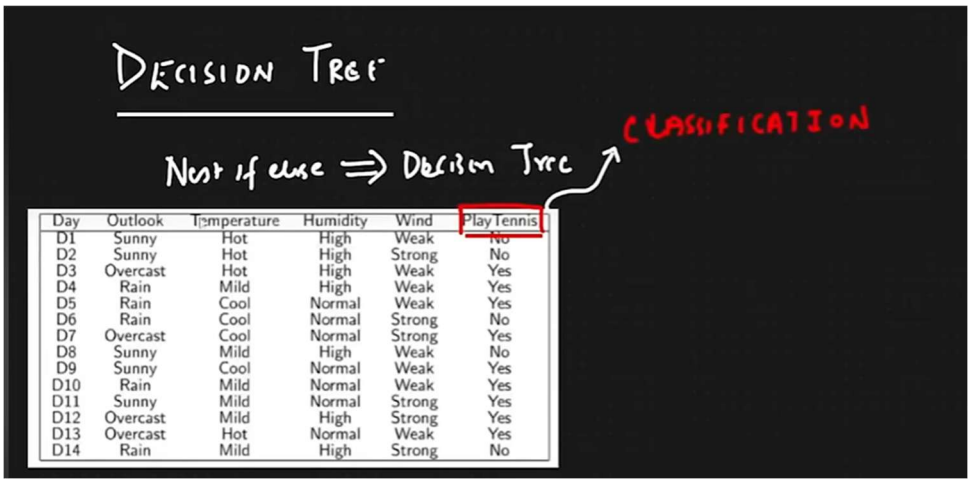
We can make a decision tree of this




Can we solve a regressrion and classification problem using these kind of nodes, whenever we talk about decision trees, they are basically using nested if else condition.

What type of Math do we normally use for this ?

For example for classification,lets take this as our problem statement :




| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

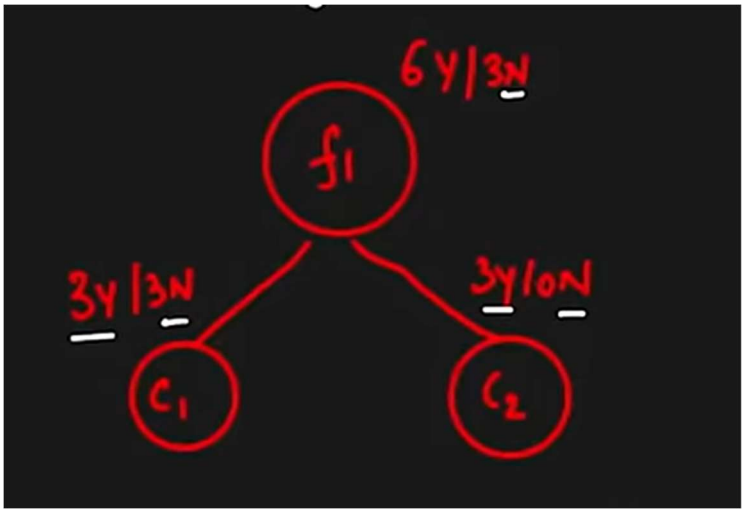
My model should predict is the persong going to play tennis or not given the condition.

- We try to get a leaf node which is completely pure, then we will stop splitting

How do we calculate the putrity, and how dow we know it is a pure split

- We use two things to determine purity
  - Entropy
  - Ginnie Coefficient/Impurity
- How the topics are selected
  - Information Gain

How does Entropy and Ginnie work:  ( By Default Classification uses Ginnie impurity)

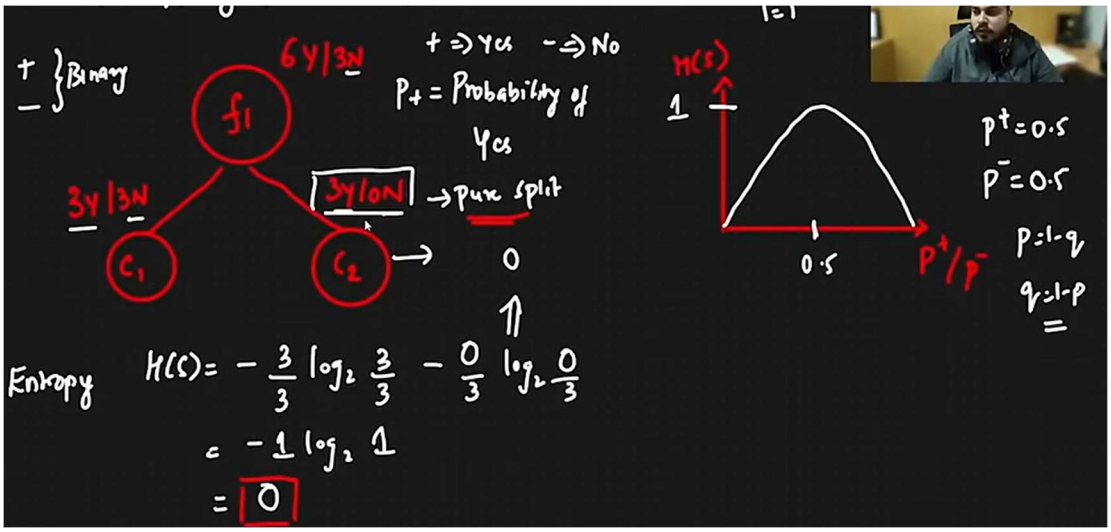

IF  we add it up it will br 6

Right side entropy

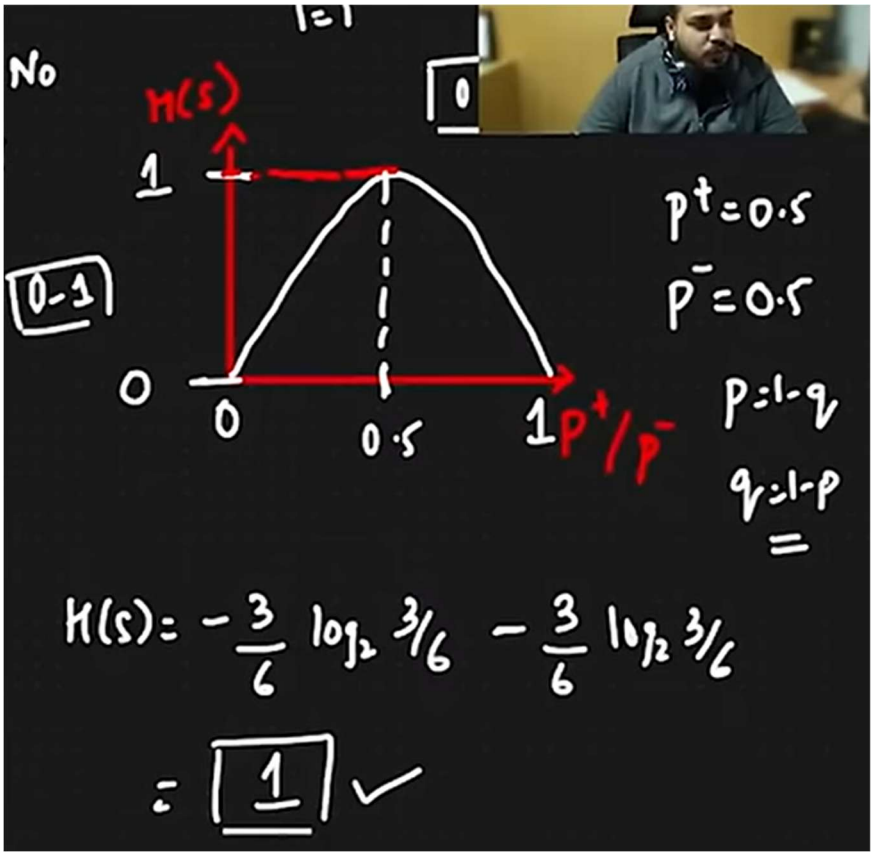
Entropy H(s) = - 3/3log_2 3/3 – 0/3log2 0/3

= -1log_2 1

= 0




Left side




The purity test is done with entropy, If we get 1 that is the impure split and if it is 0 that is a pure split/pure leaf node

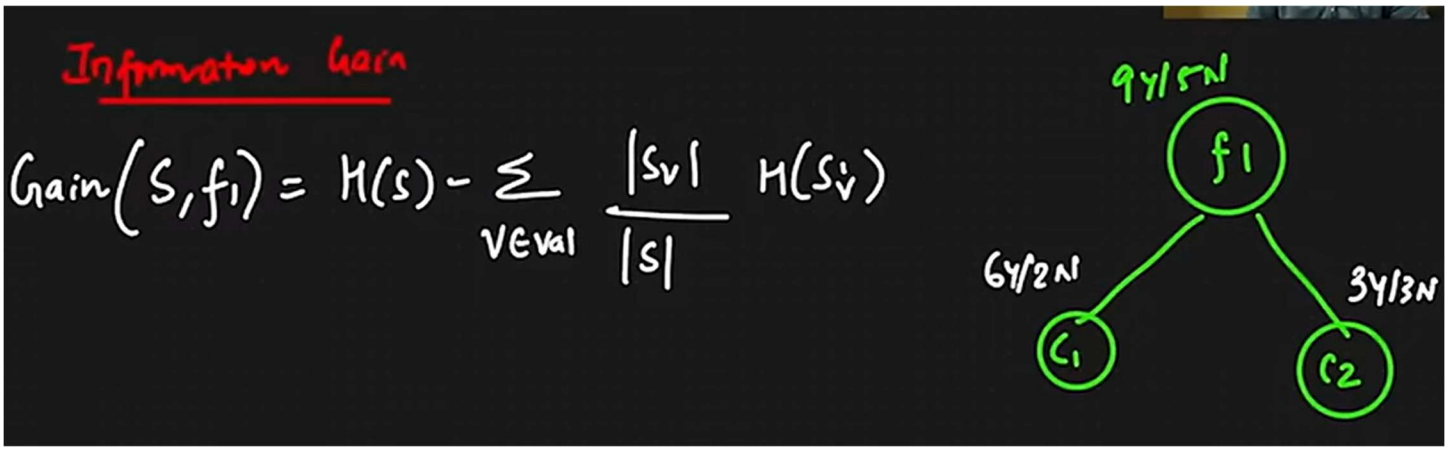
If it is impure we go further to determine the specific purity

2) How do we consider we feature to take and split ?

My ans: we will probably do an information gain process

How do we decide  that I shoyld F1 first/ F2 first/F3 first:

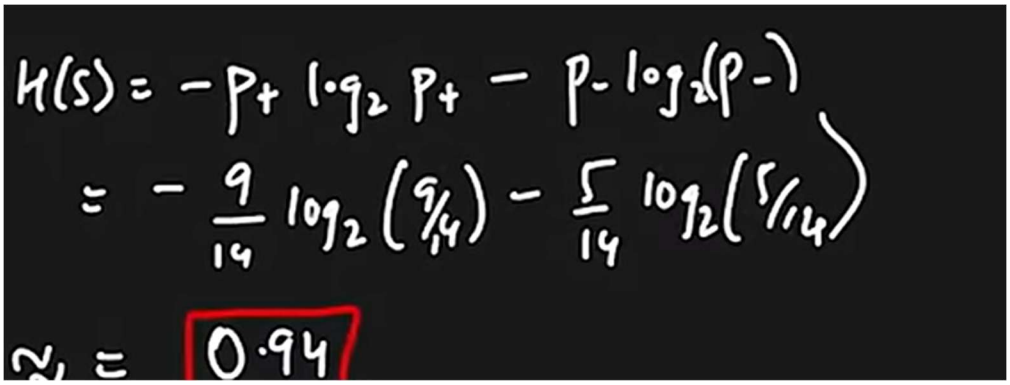
**Information Gain**




We will try to understand what is H(s), H(s) is nothing but entropy of the root node,

H(s) = -p(+)log_2(p+) – p(-)log_2p(-)

We will calculate the entropy of the root node first




Now we calculate the entropy of the C1 and C2 categories

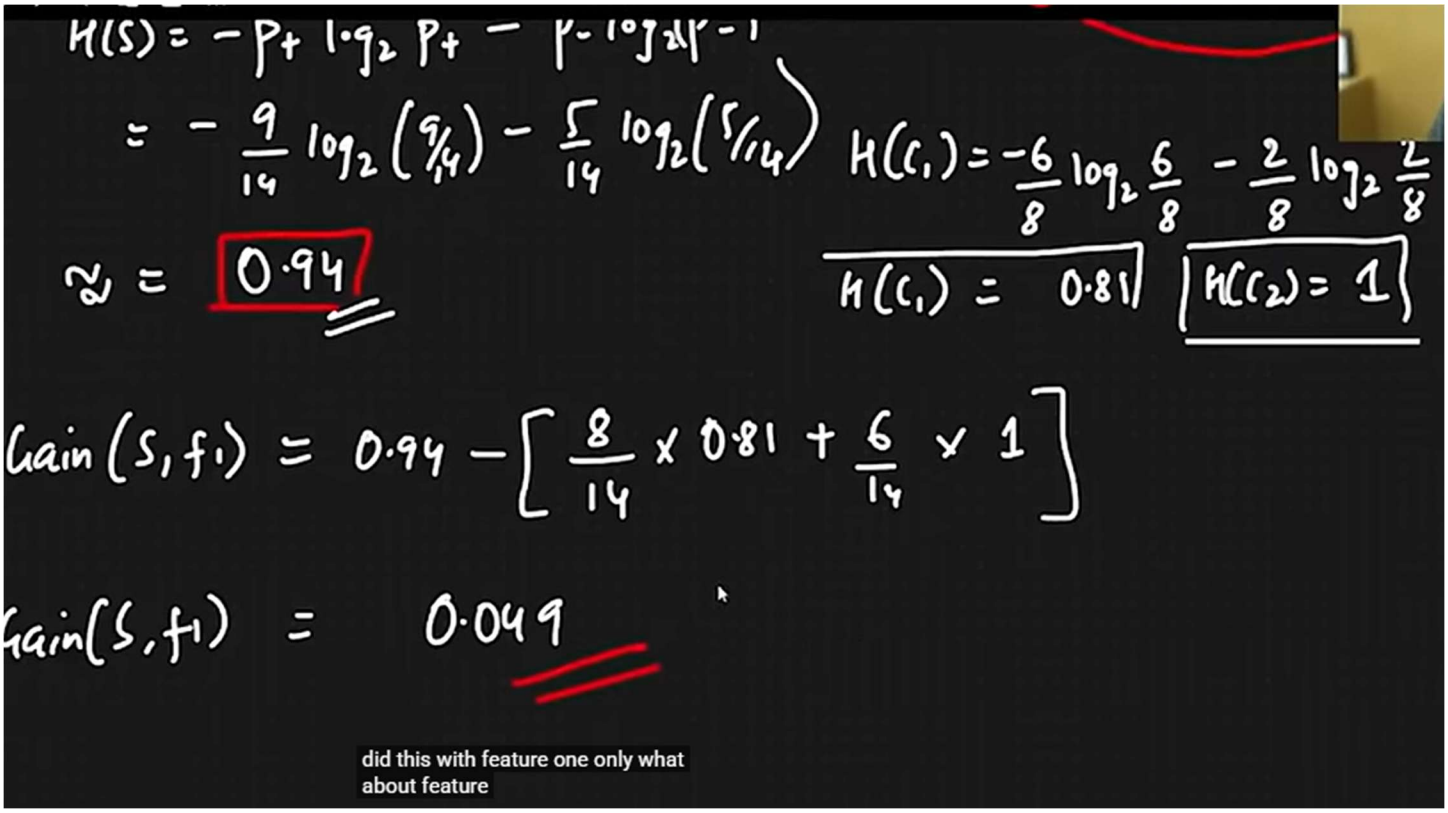
HC1 = 0.81

HC2 =  1

Sv: The total number of sample in the category for C1 it is 8,  C2 is 6

S: The total number of sample from the root C1 is 14, for C2 it is 14 (It is the same because it is coming from the same root node)
This calculation is for feature one, now what about feature 2 ?

Gain(S,F1)=

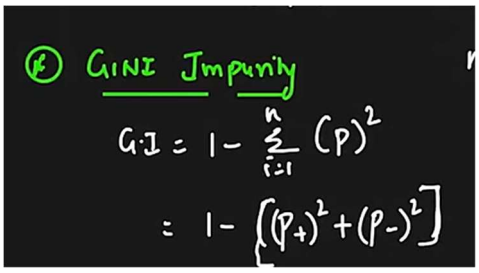


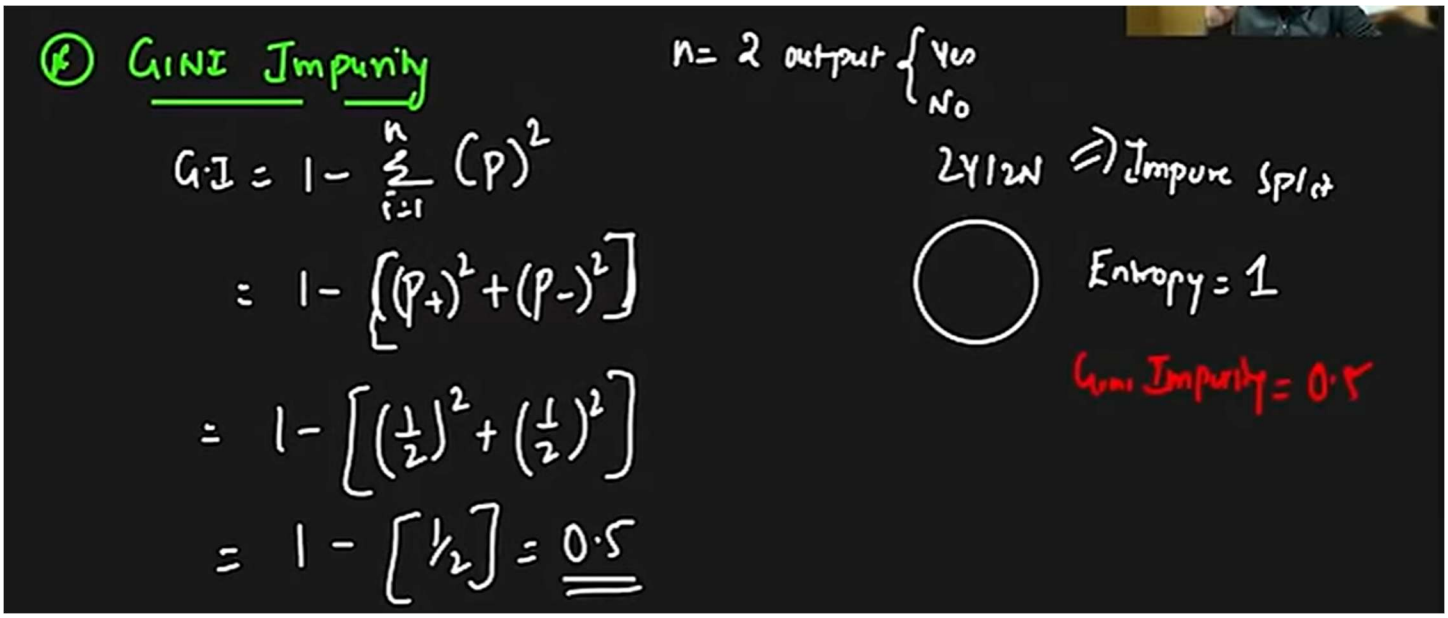

Example for feature 2 we get gain(S,F2) = 0.051

We know that Gain(S,F2)>Gain(S,F1)

So we select feature 2

Gini Impurity






When should we use ginny and when to use the other ?
Which would take more time to execute ?

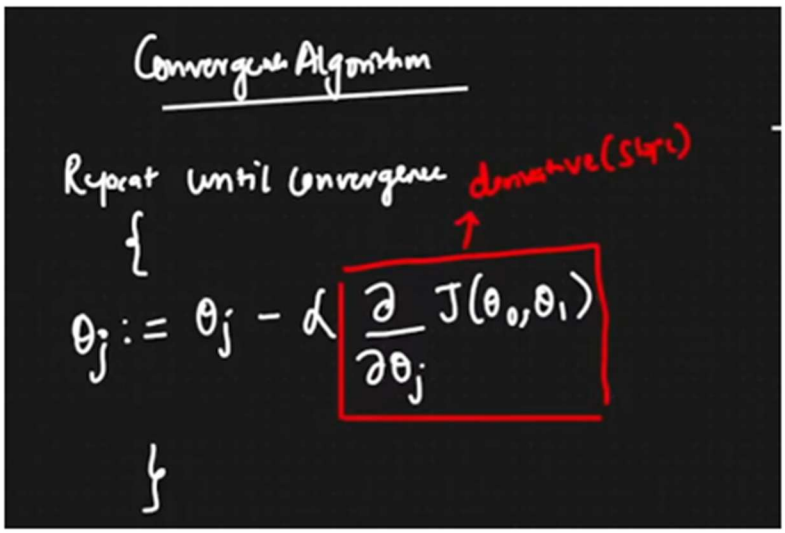
Entropy is used for small set of feature, and ginny impurity is more faster than Ginny Impurity. Hence we use Ginny impurity for large sets of data. As ginni uses simple math...

**QUESTIONS:**

# TOP 10 MOST COMMON INTERVIEW QUESTIONS

1. Tell us a bit about yourself?
2. What are your strengths & weaknesses?
3. Why do you want to work here?
4. Where do you see yourself in 5 years?
5. Why should we hire you?
6. Challenge at work & how you dealt with it?
7. Greatest professional achievement?
8. Why are you leaving your current job?
9. How do you handle stress & pressure?
10. Do you have any questions for us?

global minimum




How does it find itself ?
if we find the point in right hand side of the parabola or the right side of the parabola the formula will adjust itself relatively and also if the point is on the left side it will adjust accordingly .

**Questions Like how is Logistic Regression different from linear regression ?**

When should we use ginny and when to use the other