Excel = “Chandoo.org”

Statistic = “Khan academy

***Why do you become a data scientist?***

***To utilize my course skill knowledge and strength I want to become a data scienctist***

Normal Distribution: if probability between -1 to +1 then its 68%, if probability between -2 to +2 then its 95%, if probability between -3 to +3 then its 99.7%

Data Science:

*Data Science is not just about statistic, Machine learning and python but how well we are creating a great solution with the existing skill or with the data available with us.*

*Normalization : (Xi - min / max - min) It will normalise the data between (0,1) Range*

*Standard Scaler :(X(bar)-mean / sigma(SD))*

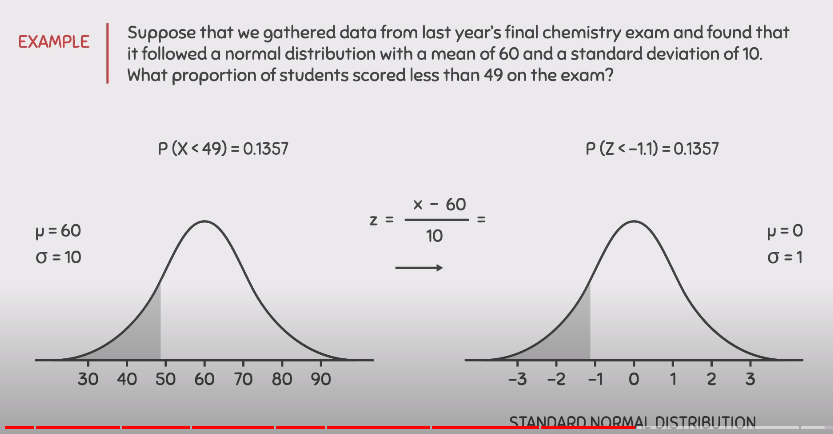
*Stander deviation : root(summation((Xi – X(bar))square / n – 1))*

*Variance : (summation(Xi – X(bar)squar / n – 1))*

*Total error: (actual – prediction)squr*

*If we apply Normal distribution all the range will become a (0,1) range*

*And In Standard Scaler al range will become a (-2 to +2) because its cover 95% of data between this range*

**

**What is mean by Random variable?**

Ans: A thing that which we cannot be predict like tomorrow weather or if we toss a coin then we will get head or tail

**Overfitting:** If a point is perfectly fitted in a line so that arises a Overfitting problem

* Eg. Suppose we told to model that this is a ball sphere is 7 cm , we can’t eat it , we can play it and many more features we are giving to the model so model will test all things and it will test also the sphere of ball its 7 cm so a thing which is 7 cm sphere so our model will predict only that thing is ball so if any ball which is having 10 cm sphere it will not predicted as a ball so that’s why overfitting problems can occur
* Solution:
* Make model simple
* Use less parameter
* Reduce features
* Reduce Noise(outliers if any)
* Dropout
* Cross-validation is a powerful preventative measure against overfitting.
* Train with more data. It won't work every time, but training with more data can help algorithms detect the signal better.
* Remove features.
* Early stopping.
* Regularization.
* Ensembling.

While we perform training and testing the data then our model will gives us accurate result but if we give any new data for testing to our model then we if will gives us wrong result it cant predict new result aapl model tech perform krel je aapn tyala train kel ahe

In overfitting our model is more promed to decision tree

**Underfitting:** If a point is not fitted in a line then we can say it is a problem of Overfitting problem

It is not able to perform relationship hence it cant predict the accurate result on testing the data

Eg. Suppose we told to model that this is a ball so just predict that is anything you you find out its sphere then its ball, so it’s a problem of underfitting our model will predict orange also as ball because we are giving very less features to our model so anything which is sphere will predict it as a ball

In underfitting our model is more promed to linear regression

Solution:

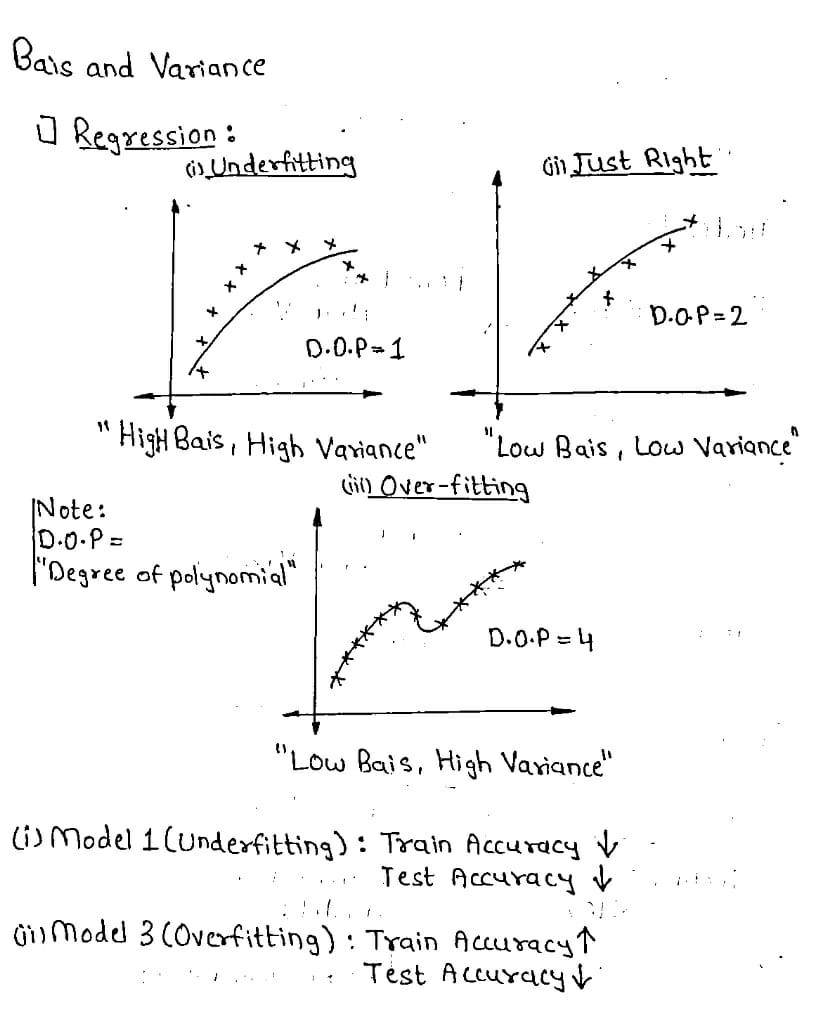
* Powerful model
* More parameters
* Use correlated feature
* Reduce the constraints on the model
* Get more data

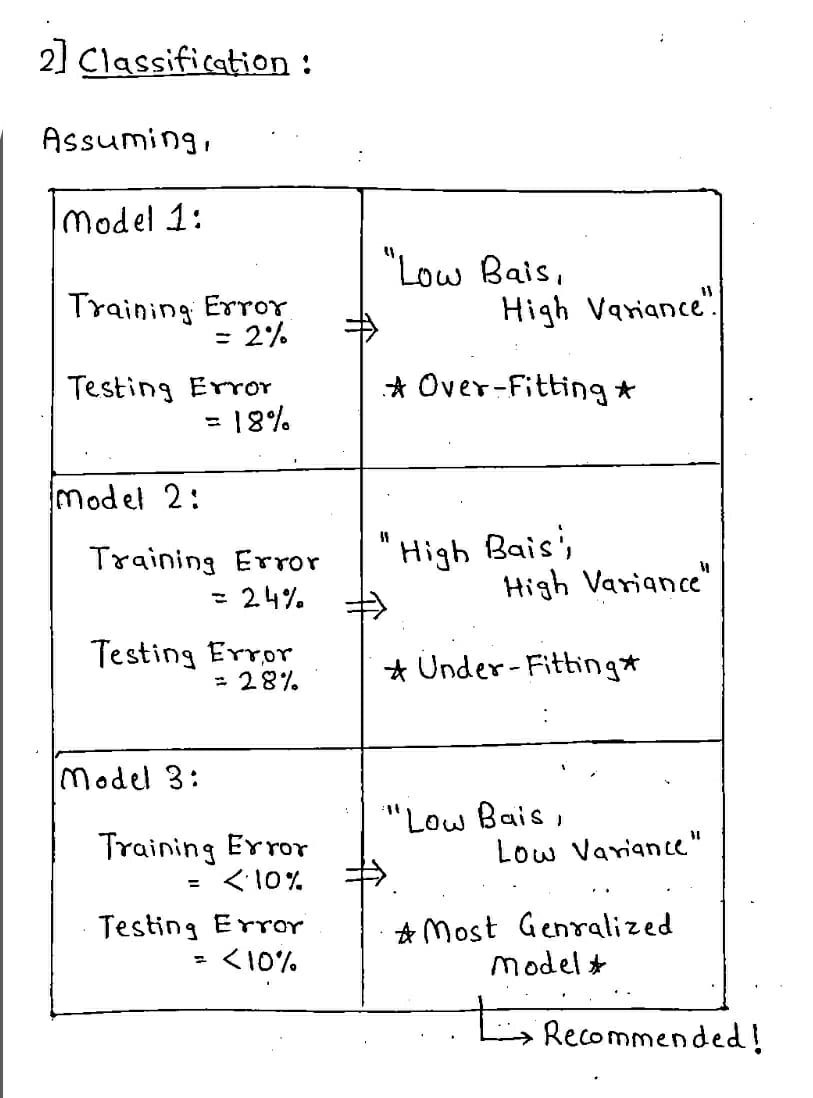
Bias and Variance

While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias.

Bias is Error resulted from Training set, while Variance is error resulted from Test set

variance means when model is take into the count the all the flacuation in the data is known as variance

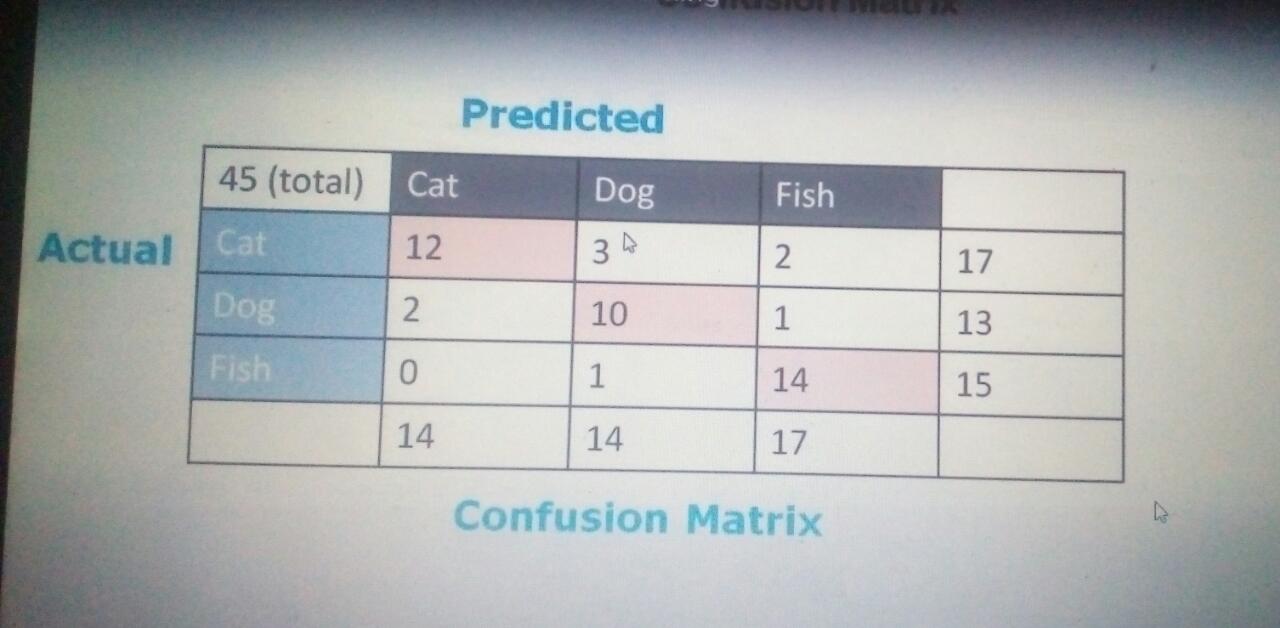


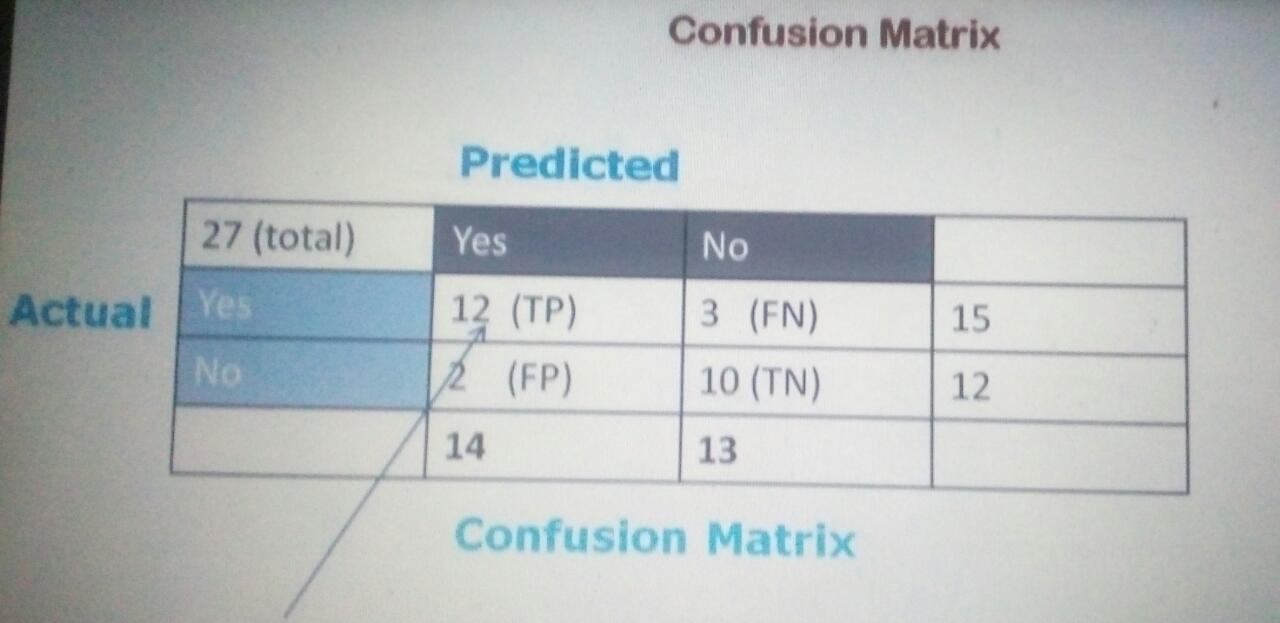


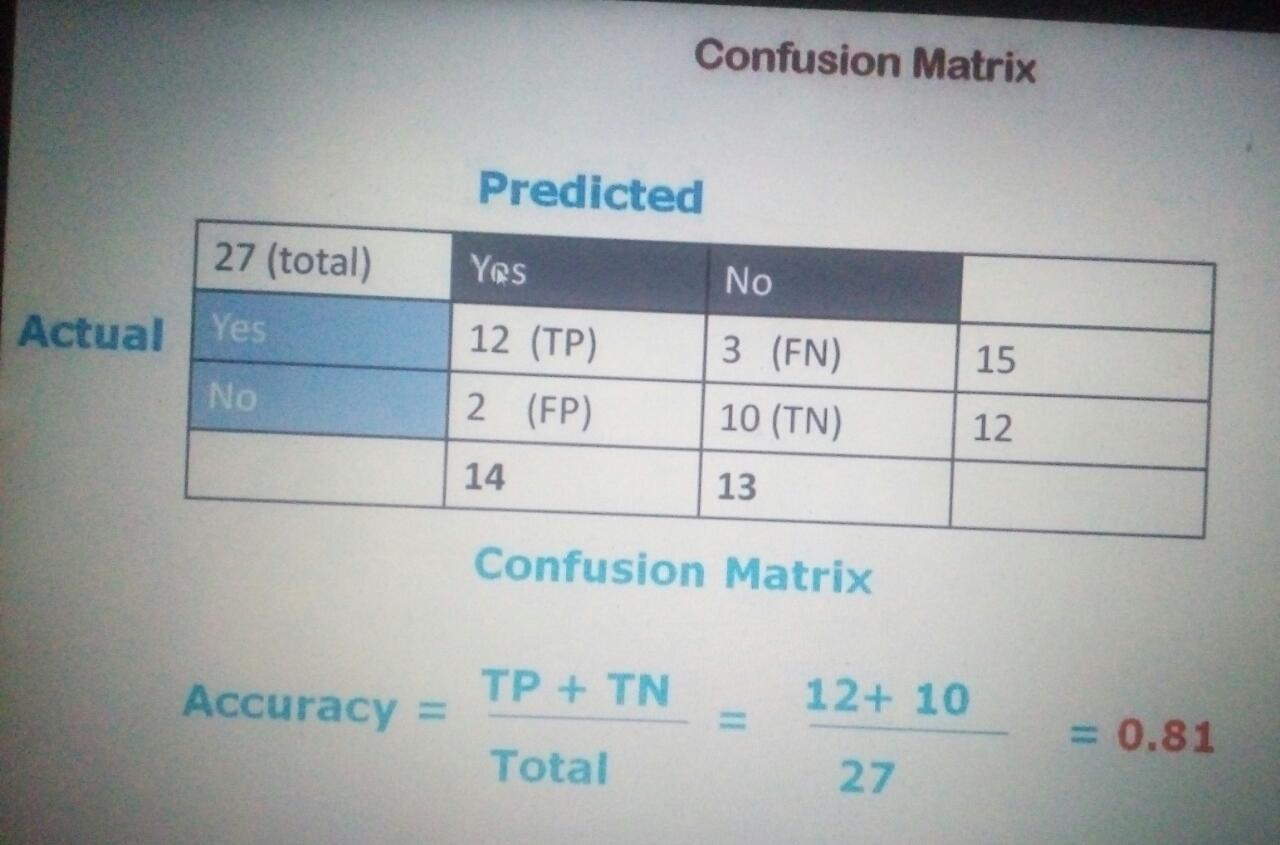
**Confusion Matrix**

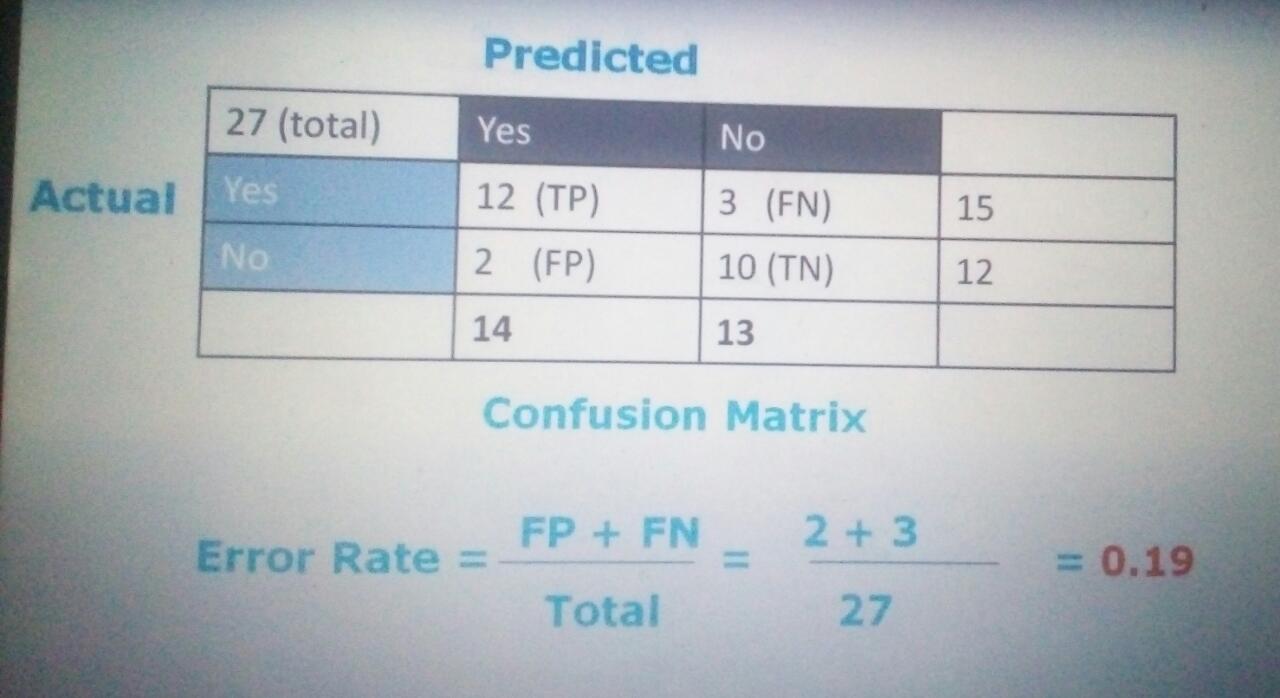
It is a matrix which is having 2 rows and 2 columns, It has 4 output that binary classification model is provide, it is used to define various measure like specificity, precision ,recall , error rate,accuracy

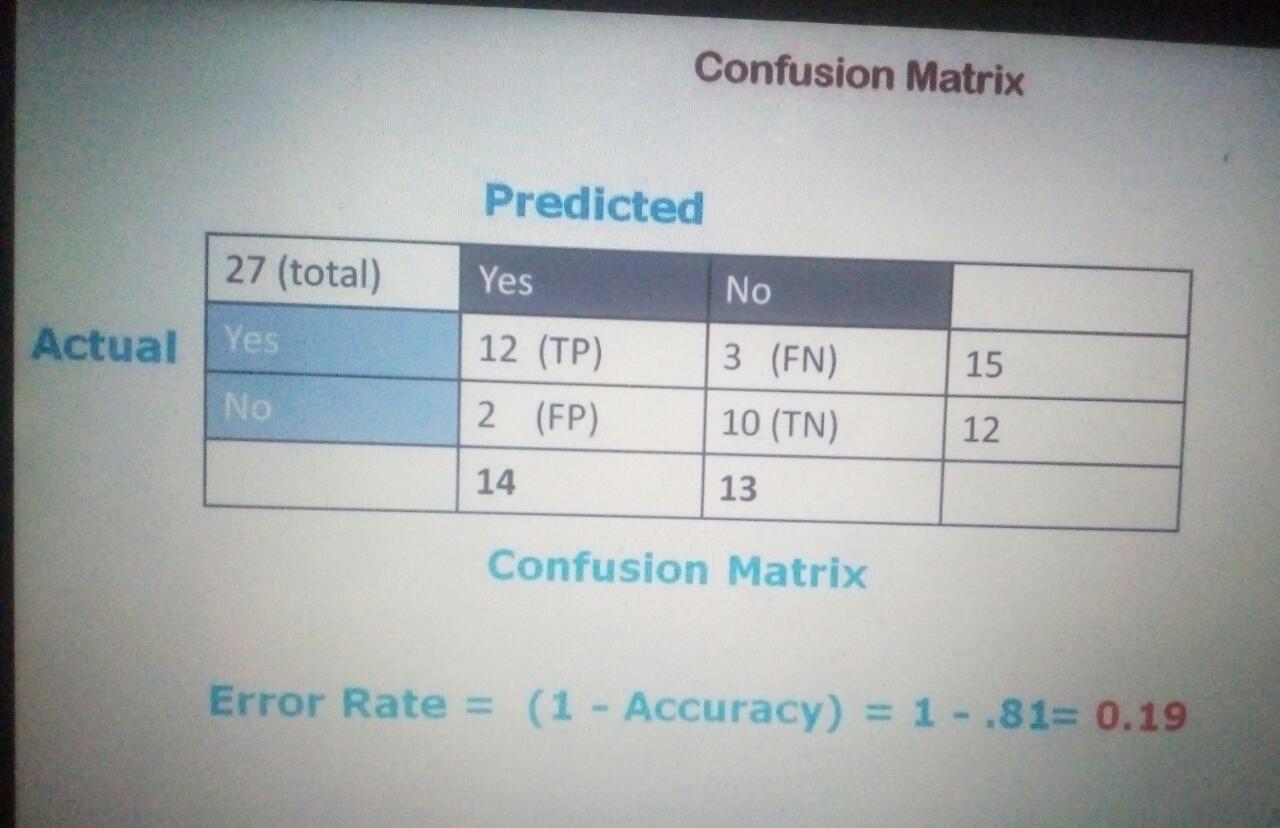
A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those the predicted by the machine learning model.The rows represent the predicted values of the target variable.

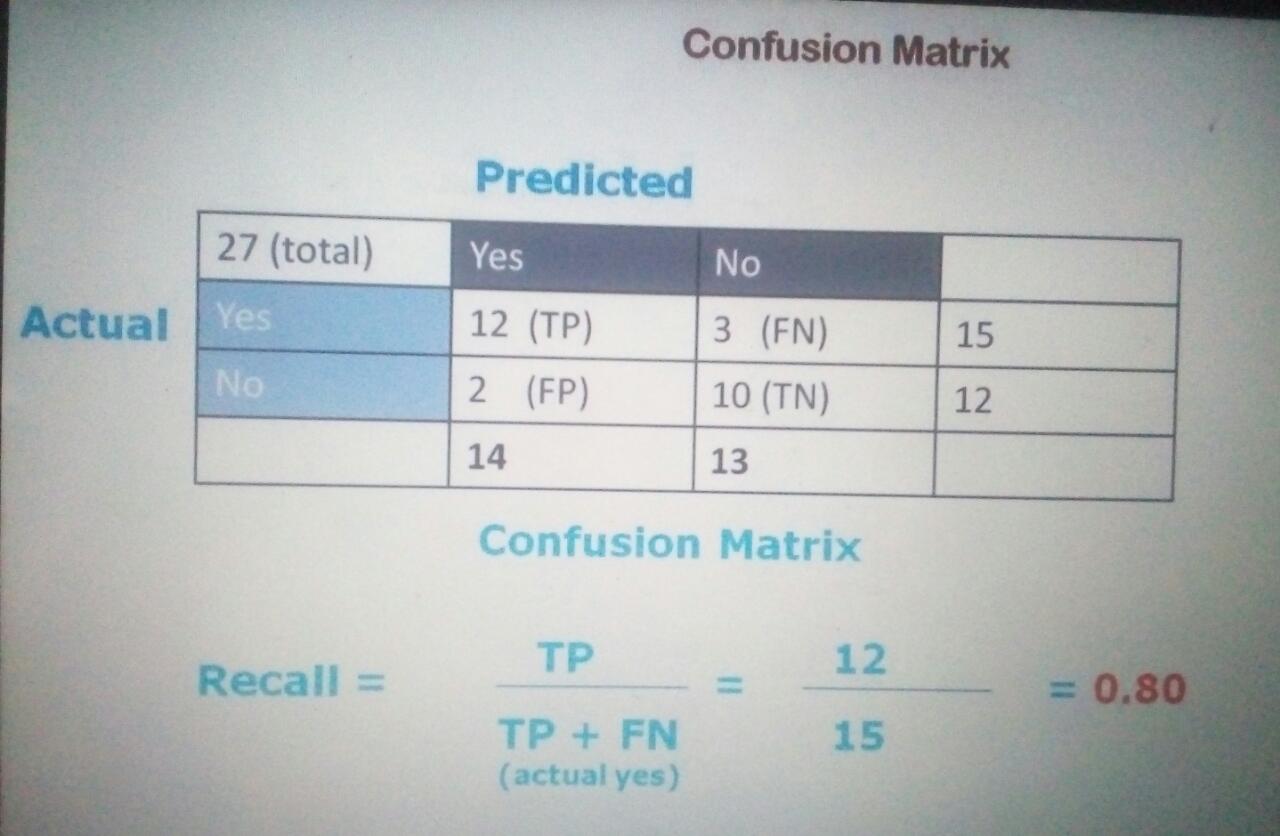
****

****

****

****

****

****

**F1 SCore : 2\*precision\*recall/precision\*recall**

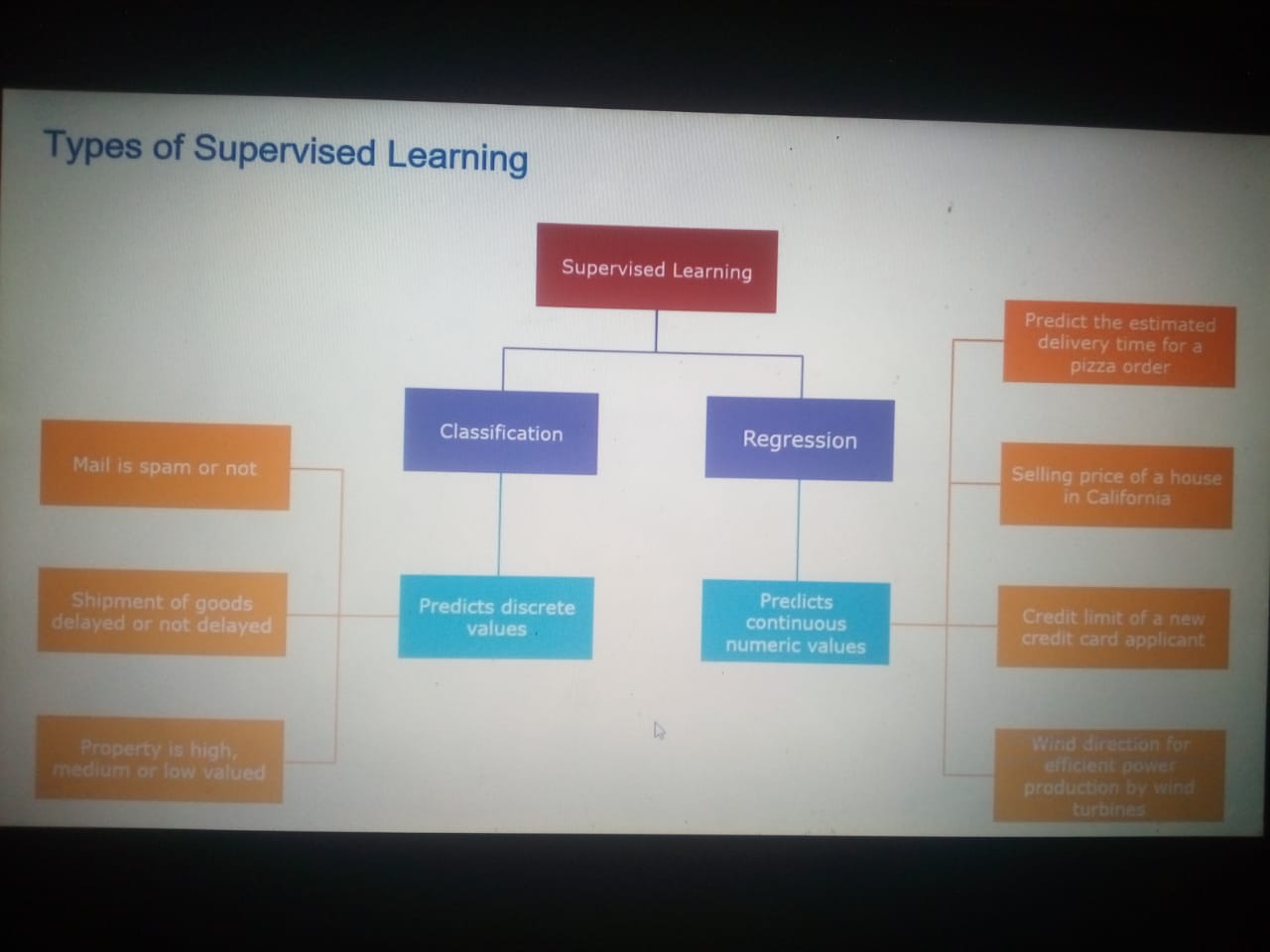
Supervised Machine Learning

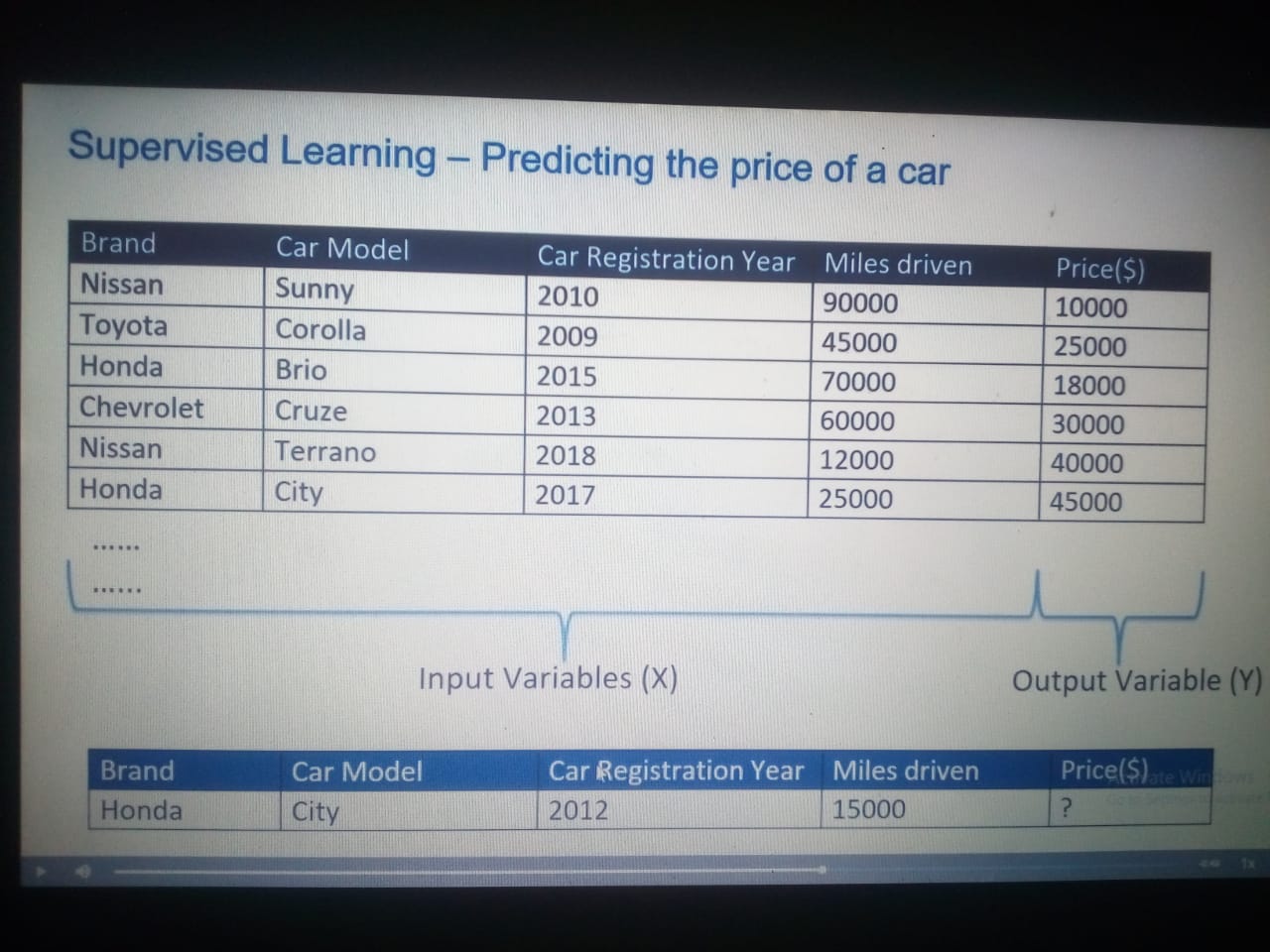
Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

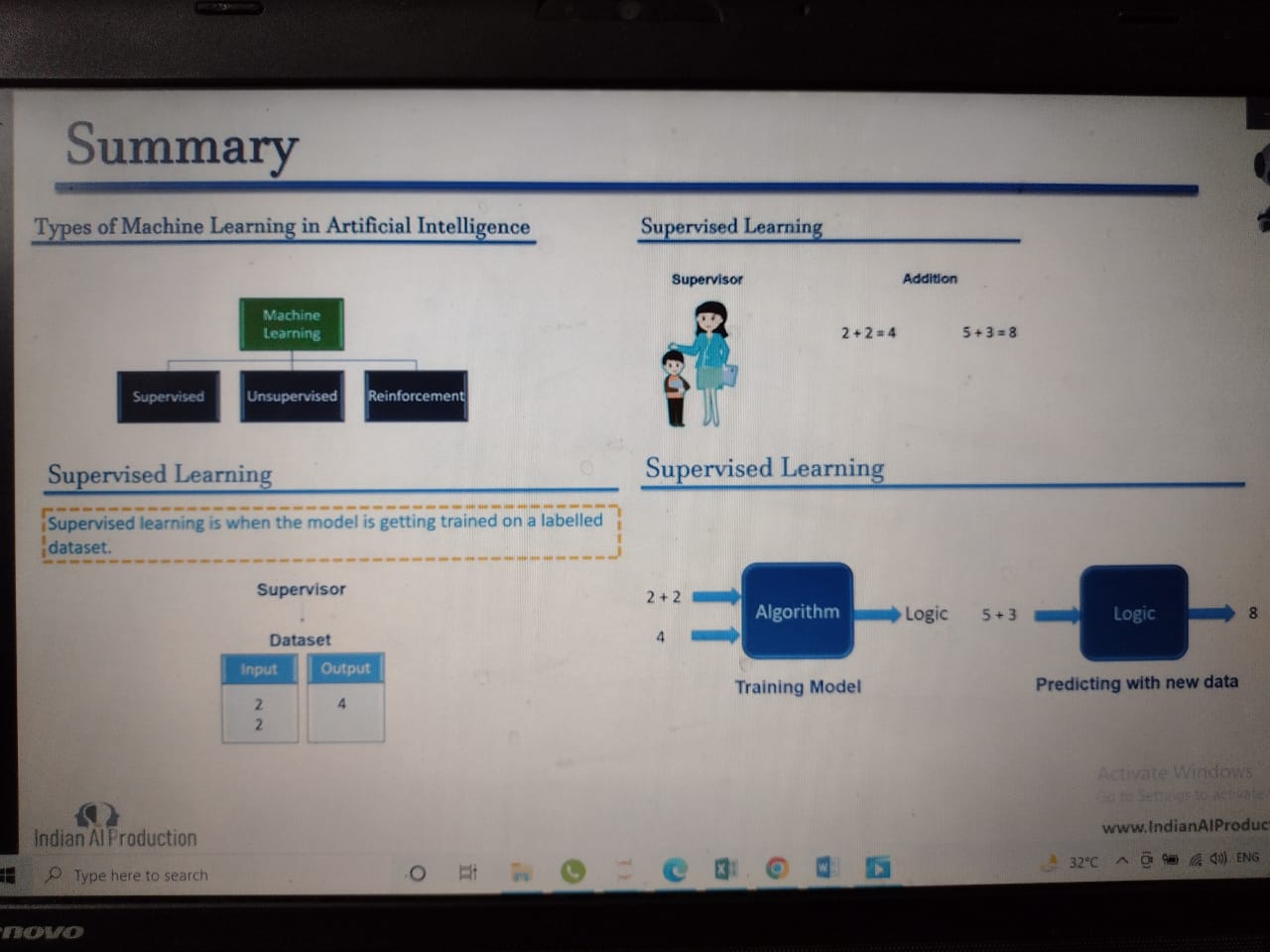
In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc







Types of supervised Machine Learning Algorithms:

1. Regression
2. Classification

**1. Regression**

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

* Linear Regression
* Multiple Linear Regression
* Polynomial Regression
* Support Vector Regression
* Decision Tree Regression
* Random forest Regression
* Logistic Regression

**Example:** Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

**2.Classification**

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

* Random Forest Classification
* Decision Trees Classification
* Linear Classifiers
* K-NN (k nearest neighbors)
* Support vector Machines
* Naïve Bayes

**Example:** The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

Advantages of Supervised learning:

* With the help of supervised learning, the model can predict the output on the basis of prior experiences.
* In supervised learning, we can have an exact idea about the classes of objects.
* Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

Disadvantages of supervised learning:

* Supervised learning models are not suitable for handling the complex tasks.
* Supervised learning cannot predict the correct output if the test data is different from the training dataset.
* Training required lots of computation times.
* In supervised learning, we need enough knowledge about the classes of object

Unsupervised Machine Learning

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.

**Example:** Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images

Why use Unsupervised Learning?

Below are some main reasons which describe the importance of Unsupervised Learning:

* Unsupervised learning is helpful for finding useful insights from the data.
* Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
* Unsupervised learning works on unlabelled and uncategorized data which make unsupervised learning more important.
* In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.
* Types of Unsupervised Learning Algorithm:
  1. Clustering
  2. Association
* **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.
* **Association**: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Unsupervised Learning algorithms:

* **K-means clustering**
* **Hierarchal clustering**
* **Principle Component Analysis**
* **Independent Component Analysis**
* **Singular value decomposition**

Advantages of Unsupervised Learning

* Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labelled input data.
* Unsupervised learning is preferable as it is easy to get unlabelled data in comparison to labelled data.

Disadvantages of Unsupervised Learning

* Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
* The result of the unsupervised learning algorithm might be less accurate as input data is not labelled, and algorithms do not know the exact output in advance.

**DAY 1.**

**DATA PREPROCESSING**

**Step 1 . Importing the required Libraries**

Pandas and Numpy are the essential libraries which we used every time

* Numpy is used for mathematical functions
* Pandas is used for import and manage the data set

**Step 2. Importing the Data set**

Data set are generally available in Pandas library in which we used csv file

**Step 3. Handling the missing data**

Replace the null data with mean or median

Mean: data=data.fillna(data.mean())

Median: data=data.fillna(data.median())

Standard Deviation: data=data.fillna(data.std())

from sklearn.impute import SimpleImputer

>>> imp = SimpleImputer(missing\_values=np.nan, strategy='mean')

**Step 4 Encoding catgorical Data**

Categorical data contains in label format like Yes or No.so we have to convert that data in binary form so we used LabelEncoder class from sklearn.preprocessing library

(1) from sklearn.preprocessing import OrdinalEncoder

ord\_enc = OrdinalEncoder()

dataframe["size\_category"] = ord\_enc.fit\_transform(dataframe[["size\_category"]])

(2) from sklearn import preprocessing

# label\_encoder object knows how to understand word labels.

label\_encoder = preprocessing.LabelEncoder()

# Encode labels in column 'species'.

df['species']= label\_encoder.fit\_transform(df['species'])

df['species'].unique()

**Step 5 Splitting the dataset into Test Set and Training set**

we make to partitions of dataset one for training themodel called as training set

and another is trained model called as test set,its generally 80/20 .We used train\_test\_split()

method of sklearn.crossvalidation library.

>>> import numpy as np

>>> from sklearn.cross\_validation import train\_test\_split

>>> a, b = np.arange(10).reshape((5, 2)), range(5)

>>> a

array([[0, 1],

[2, 3],

[4, 5],

[6, 7],

[8, 9]])

>>> list(b)

[0, 1, 2, 3, 4]

>>>

>>> a\_train, a\_test, b\_train, b\_test = train\_test\_split(

... a, b, test\_size=0.33, random\_state=42)

...

>>> a\_train

array([[4, 5],

[0, 1],

[6, 7]])

>>> b\_train

[2, 0, 3]

>>> a\_test

array([[2, 3],

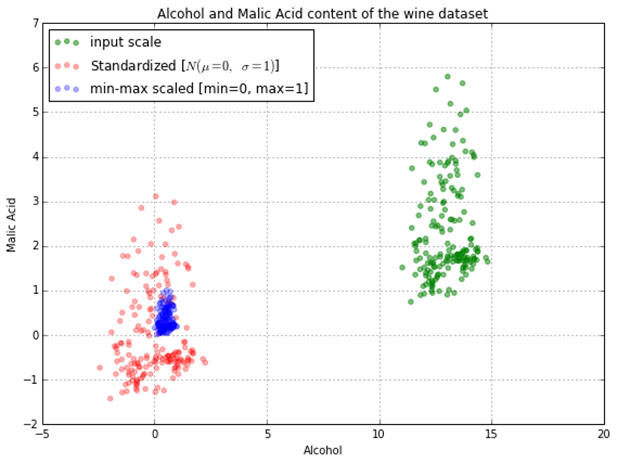
[8, 9]])

>>> b\_test

[1, 4]

**Step 6 Feature Scaling**

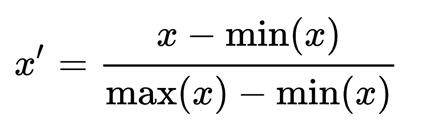
Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.ust to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling technique.



**Method of features scaling**

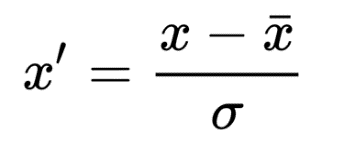
**Normalization**

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

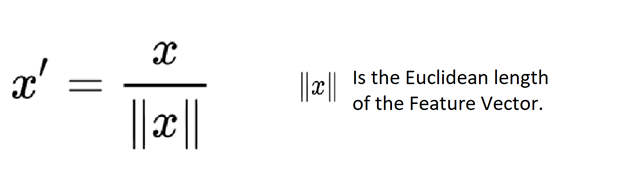
**Standardization**

Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

**Scaling to unit length**

The aim of this method is to scale the components of a feature vector such that the complete vector has length one. This usually means dividing each component by the Euclidean length of the vector:



**DAY 2.**

**Simple Linear Regression**

**Two main Objectives**

Establish if there is a relationship between to variables

eg.Income and spending,wage and gender, student height and exam score

**Forecast new observations**

can we use what we know about the relationship to forecast unobserved values?

Eg. What will are sales be over the net quarter? What will the ROI of a new store opening be contingent on store attrinutes?

**Variable Role**

**Dependent Variable:**

This is the variable whose values we want to explain or forecast

Its value depend on something else

e denoe is as Y

**Independent Variable:**

This is the variable that explains the other one

Its value are independent

We demote as X

**Equation**

y= B0+ B1x

we call it linear because the equation represents a straight line in a bidimensional plot

but there are some errors in distance beteen points from line

y= B0+ B1x+E

y is the dependent var

x is the indeendent var

B0 is the intercept or constant

B1 is x's slope or coefficient

E is the error term

**Simple Linear Regression**

simple-One independent variable and one dependent variable

Multiple Regression-Two or more independent variabable

b1=∑(xi-x(bar))(yi-y(bar))/∑(xi-x(bar))squr

b0=y(bar)-b1x(bar)

estimated regression line

y=b0+b1x

Coefficient of determination:

Ho ell does the regression line fit the data?

r(squr)=SSR/SST

where:

SSR=sum of squares due to regression=∑(y-y(bar))(squr)

SST=total sum of squares=∑(yi-y(bar))(squr)

SSE=sum of squares due to error=∑(yi-y)(squar)

SST=SSR+SSE

**Steps:**

**1. Data preprocessing**

**2. Fitting simple linear regression model to theTraining set:**

To fit the dataset into the model we will use LinearRegression class from sklearn.linear\_model library

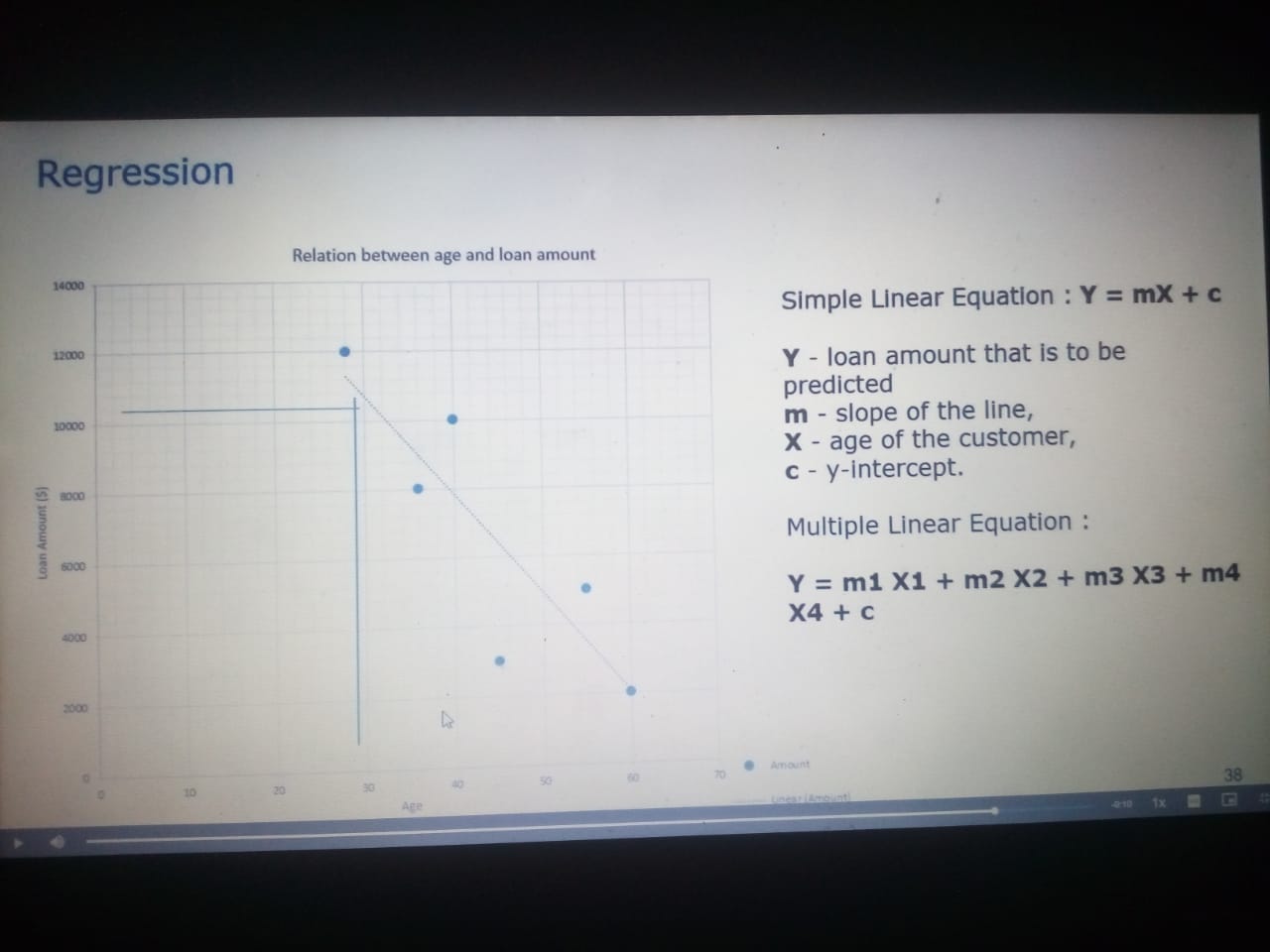
then we make an object regressor of linearRegression class.No e wil fit the regressor obj into our dataset using fit()

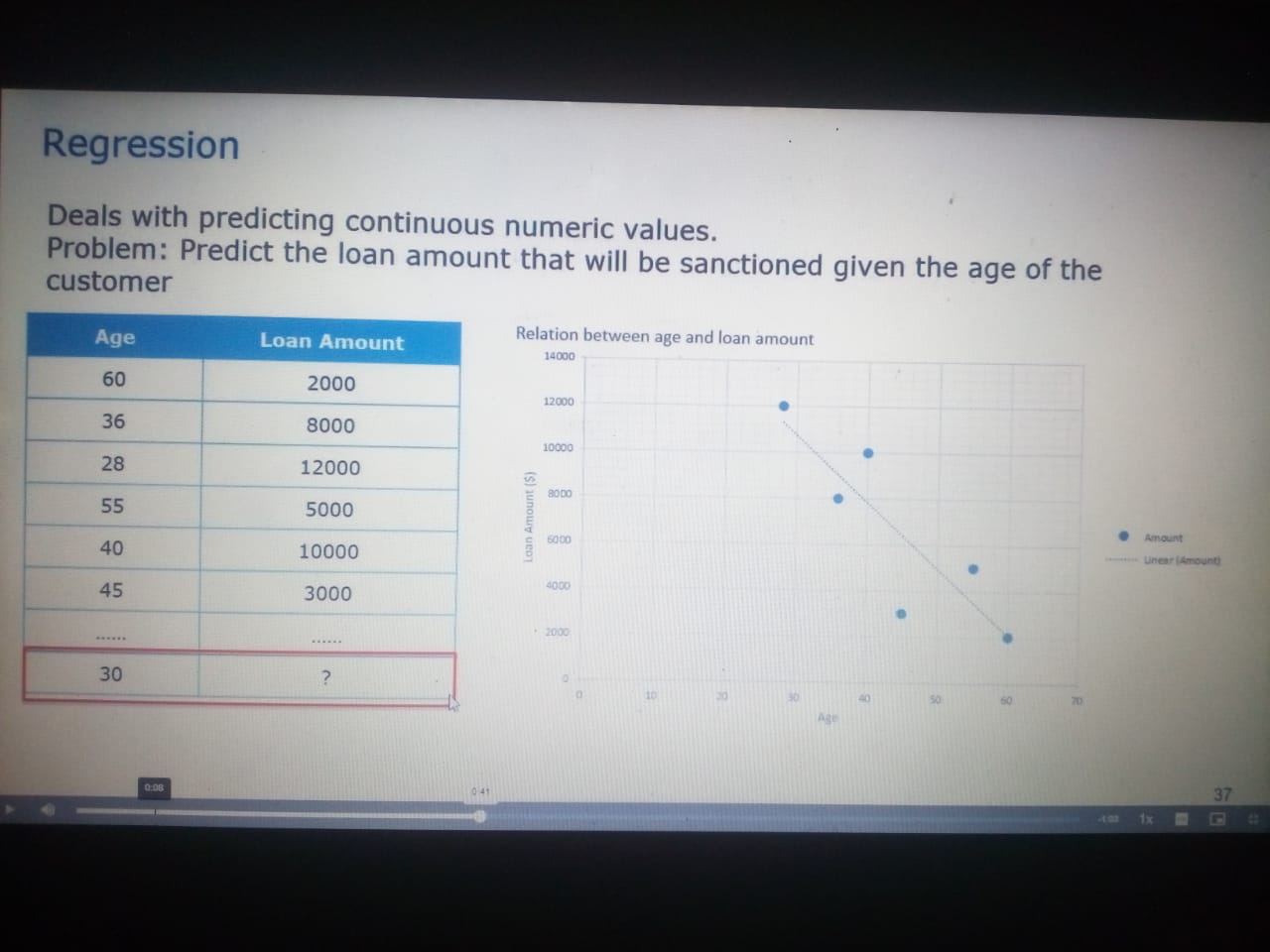
**3. Predicting the result**

we will predict the observations from our test set .the we save the output in a vector y\_pred.e use prdict method of LinearRegression Class

**4. Visualization**

We will use matplotlib.pyplot library to make scatter plots of our trating set results and test set results to see how close our model predicted the values

****

****

**DAY 3.**

**Multiple Linear Regression**

MLR attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data.it is similar to SLR only difference is that you can use it to find out which factor has the highest impact on the predicted output and how different variables relate to each other.

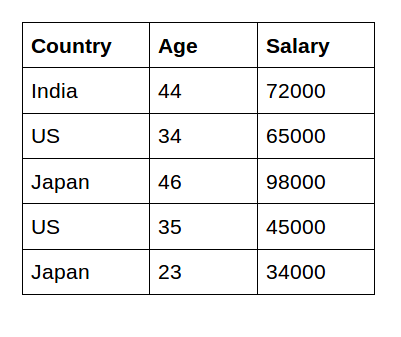
Explain y using more than one indepedent varible

y=B0+B1x1+B2x2+B3x3+E

**One-Hot Encoding vs. Label Encoding using Scikit-Learn**

**Label Encoding:**

Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.



*# Import label encoder*

*from sklearn import preprocessing*

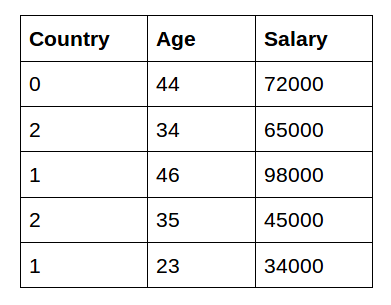
*# label\_encoder object knows how to understand word labels.*

*label\_encoder = preprocessing.LabelEncoder()*

*# Encode labels in column 'Country'.*

*data['Country']= label\_encoder.fit\_transform(data[‘Country'])*

*print(data.head())*



As you can see here, label encoding uses alphabetical ordering. Hence, India has been encoded with 0, the US with 2, and Japan with 1.

**One-Hot Encoding:**

In the above scenario, the Country names do not have an order or rank. But, when label encoding is performed, the country names are ranked based on the alphabets. Due to this, there is a very high probability that the model captures the relationship between countries such as India < Japan < the US.

One-Hot Encoding is another popular technique for treating categorical variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature.

*# importing one hot encoder*

*from sklearn from sklearn.preprocessing import OneHotEncoder*

*# creating one hot encoder object*

*onehotencoder = OneHotEncoder()*

*#reshape the 1-D country array to 2-D as fit\_transform expects 2-D and finally fit the object*

*X = onehotencoder.fit\_transform(data.Country.values.reshape(-1,1)).toarray()*

*#To add this back into the original dataframe*

*dfOneHot = pd.DataFrame(X, columns = ["Country\_"+str(int(i)) for i in range(data.shape[1])])*

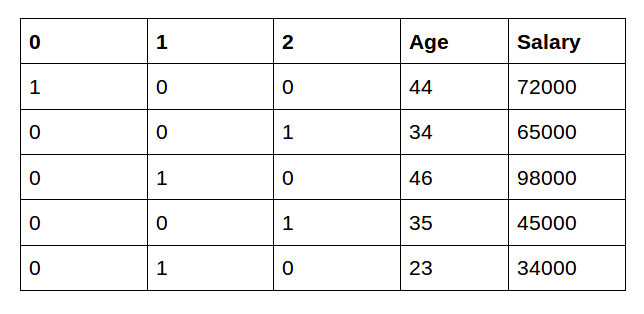
*df = pd.concat([data, dfOneHot], axis=1)*

*#droping the country column*

*df= df.drop(['Country'], axis=1)*

*#printing to verify*

*print(df.head())*



**DAY 4**

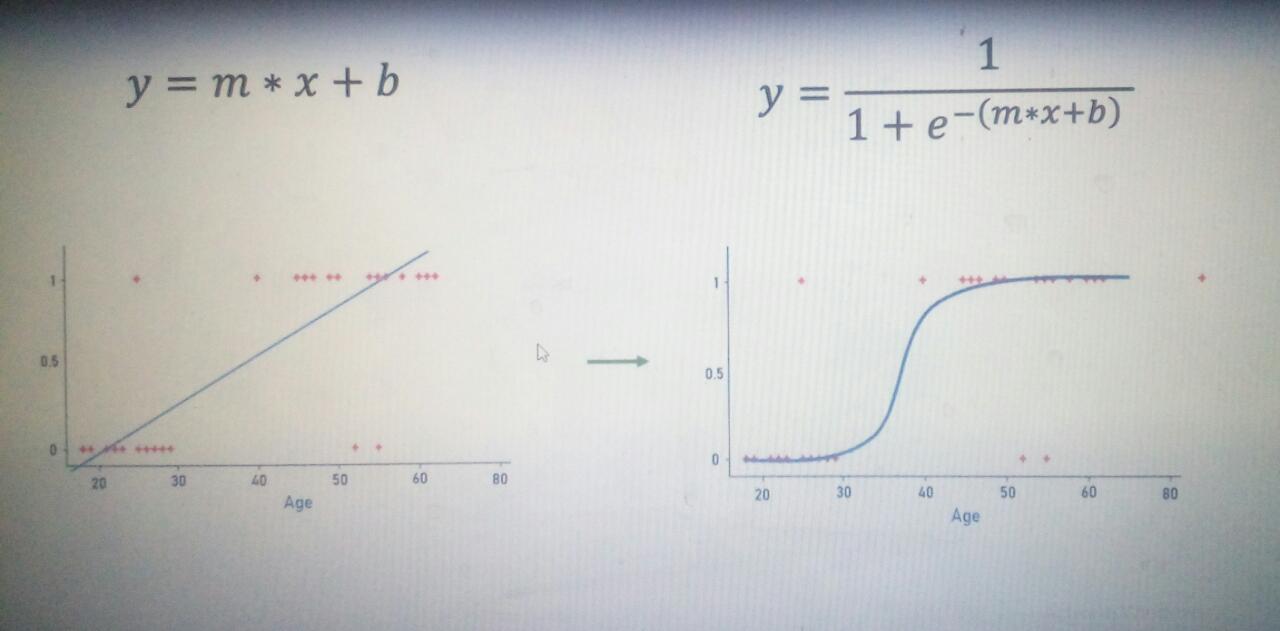
**Logistic Regression**

Logistic vs Linear:

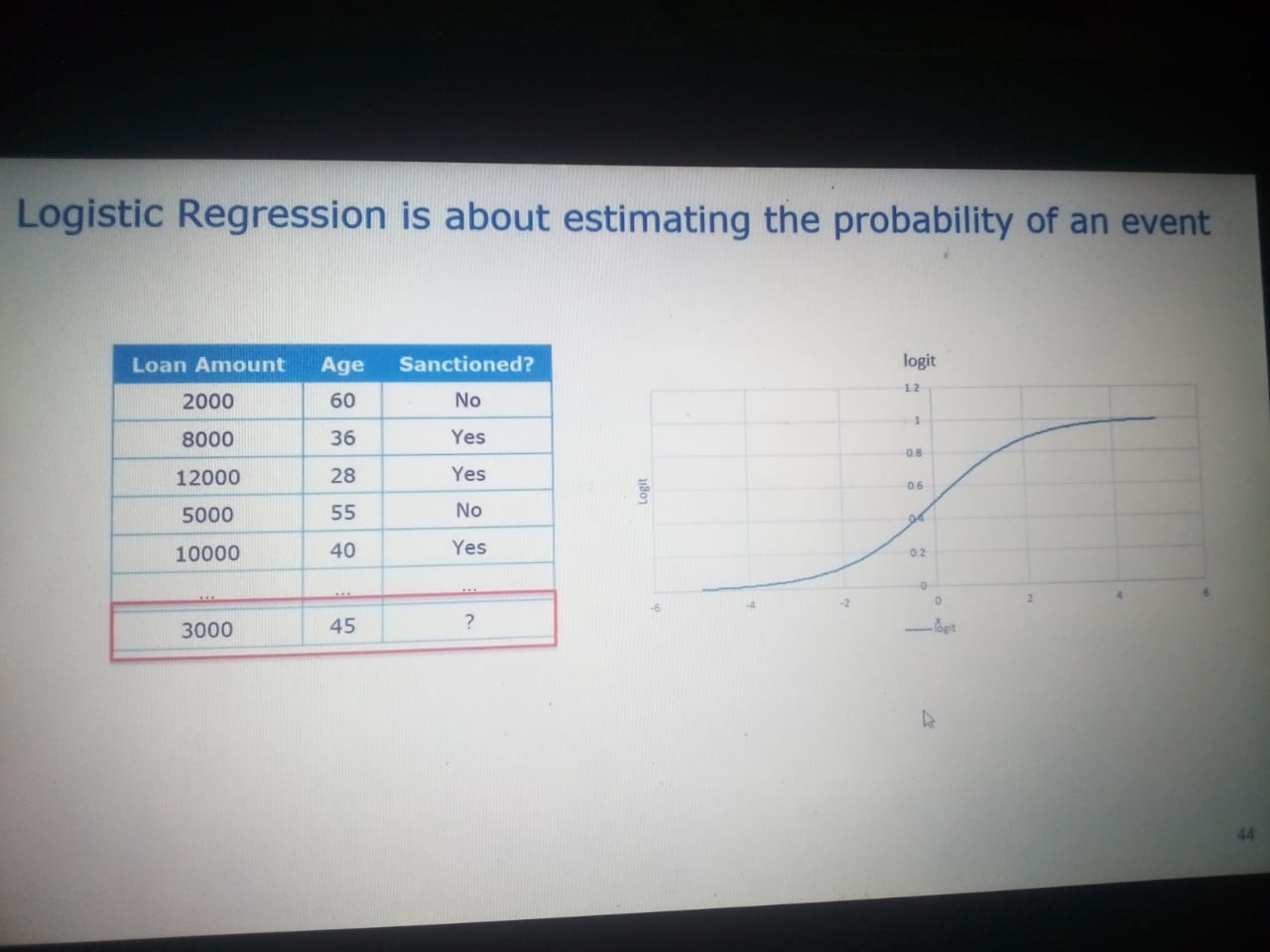
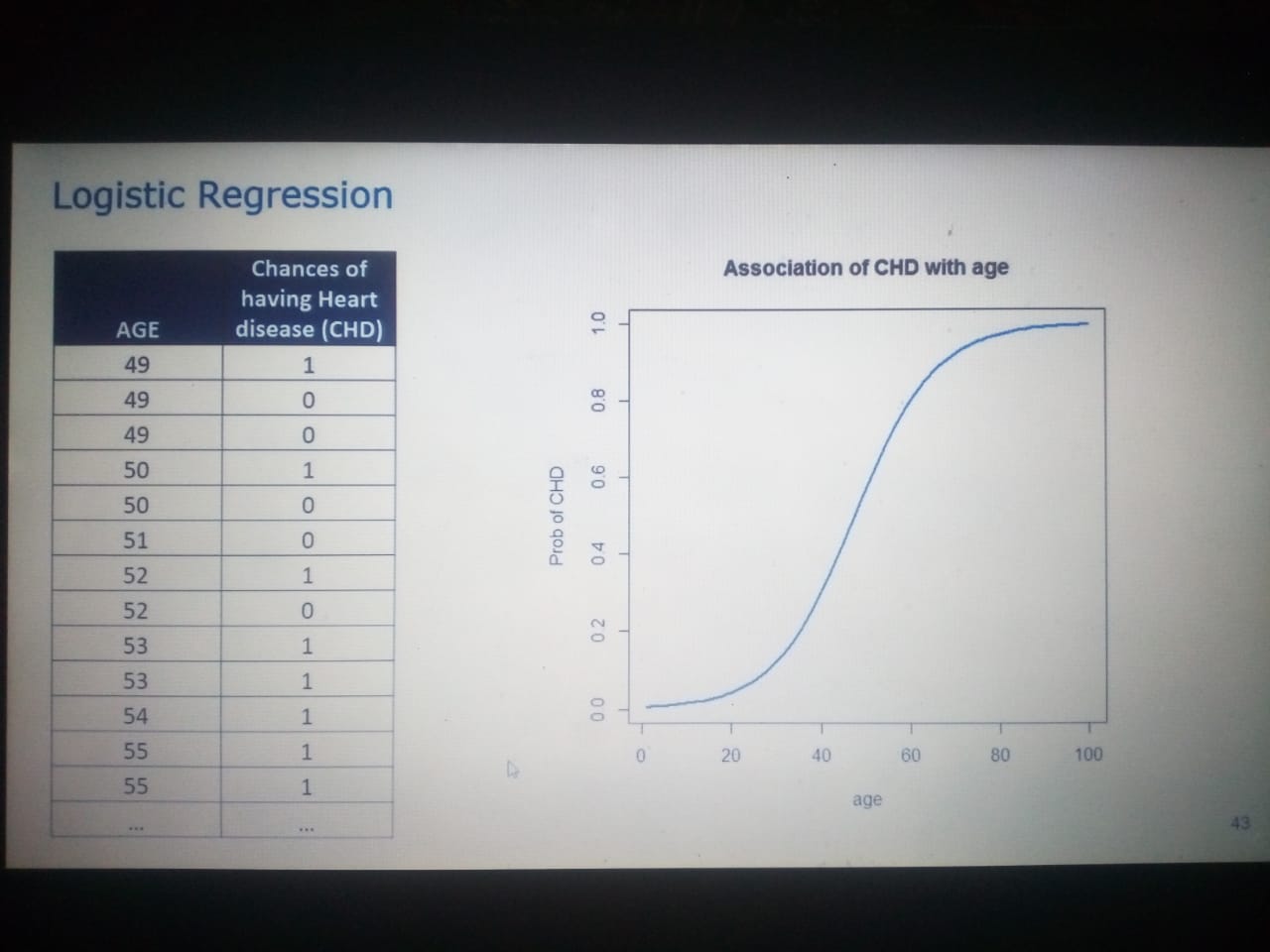
Logistic regression gives you a discrete outcome but linear regression gives a continuous outcome.

Sigmoid(z) = 1/1+e(pow)-z

It will coverts input values to 0’s or 1’s,it will create s shape line



Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems



**DAY 5**

**K Nearest Neighbours**

KNN algorithm used for both classification and regression technique

* 1. **Classification:**

**It will find out the new point suppose it got 3 blue point and 2 red point so according to the K value new point will be o that category where has max class o points**

* **Select the K value always choose odd**
* **Calculate the distance**
* **Sort in ascending order**
* **If regression, then return mean of the K labels**

**The centroid will be in middle and it will find out nearest points average on the bases of K value**

* **If classification, then return mode of the K labels**
  1. **Euclidean**

Or Euclidean distance is defined as the distance between two points. To find the distance between two points, the length of the line segment that connects the two points should be measured. real value input variables, the most popular distance measure is Euclidean distance.

**d =√[(x2– x1)2 + (y2– y1)2]**

Where,

“d” is the Euclidean distance

(x1, y1) is the coordinate of the first point

(x2, y2) is the coordinate of the second point.

KNN is also called lazy algorithm because its instance based

Euclidean distance is calculated as the square root o the sum o the square difference between a new point and an existing point across all input attributes

Other [popular distance measures include:

Hamming distance

Manhattan distance

Minkowski distance

****

**DAY 6**

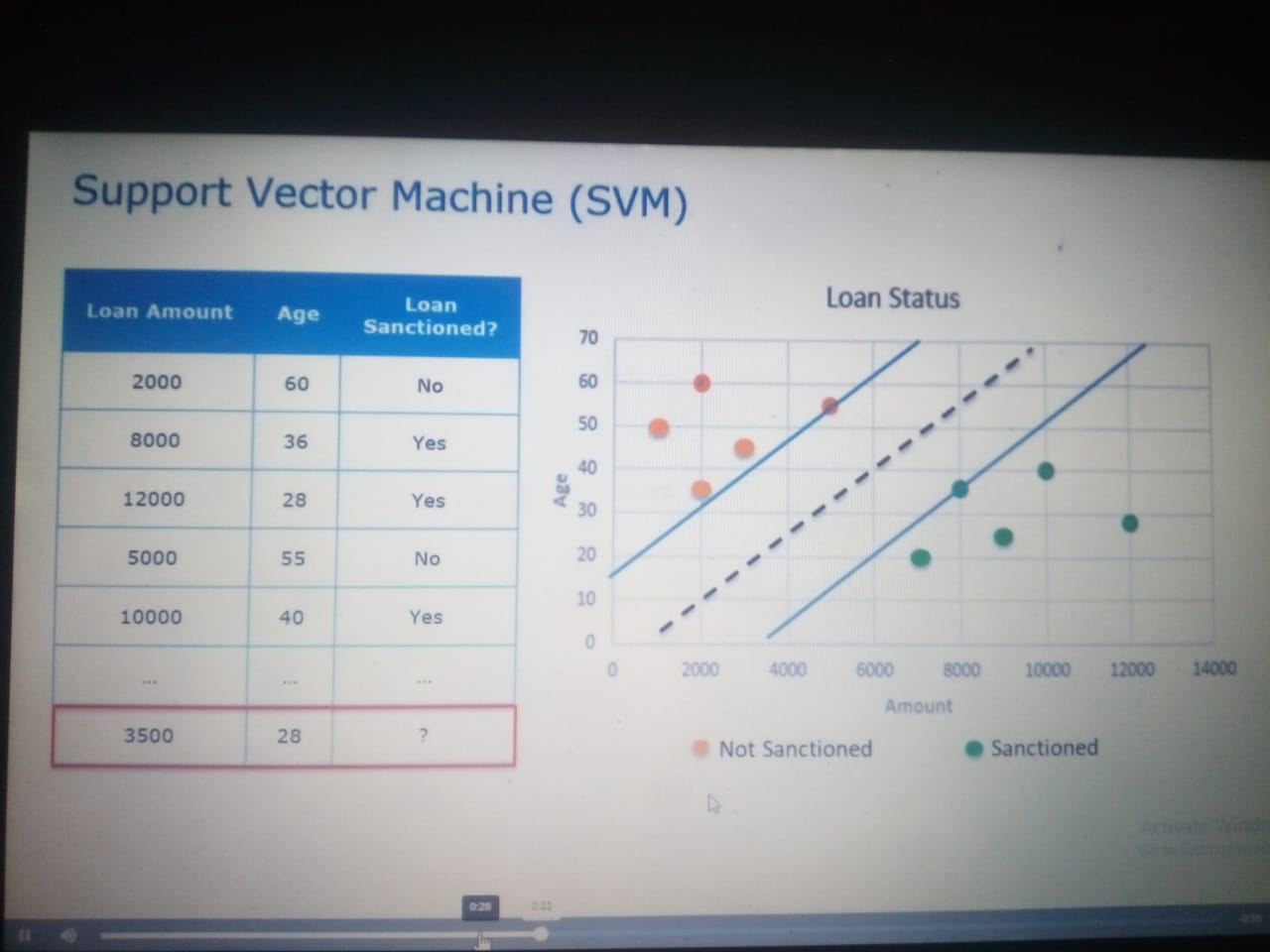
**Support Vector Machine**

SVM is a supervised machine learning algorithm which is used or both Regression and classification Technique but it is mostly used classification technique

In this two technique is used

1.Glasso

2.Regularization



**DAY 7**

**Neural Network**

A neural network is a network or circuit of biological neurons, or, in a modern sense, an artificial neural network, composed of artificial neurons or nodes.

Input Data- algorithm - Convert in Meaningful data – NN(learning algorithm) Improvement

Suppose we saw some people in office so human brain will detect that faces and next day the can easily recognize that faces but in computer language computer will scan that all the images but I some people did shave or some’s look is change then in that situation NN is improve the model and gave the best prediction

To detect a color or to detect a attribute easily we can detect because our brain is trained from childhood.

If I want to replicate them to machine so machine can made up of neurons this is known as neurons

It is very big like

Eg. Google Translate

There are two types o NN

Convolutional Neural Network(CNN):

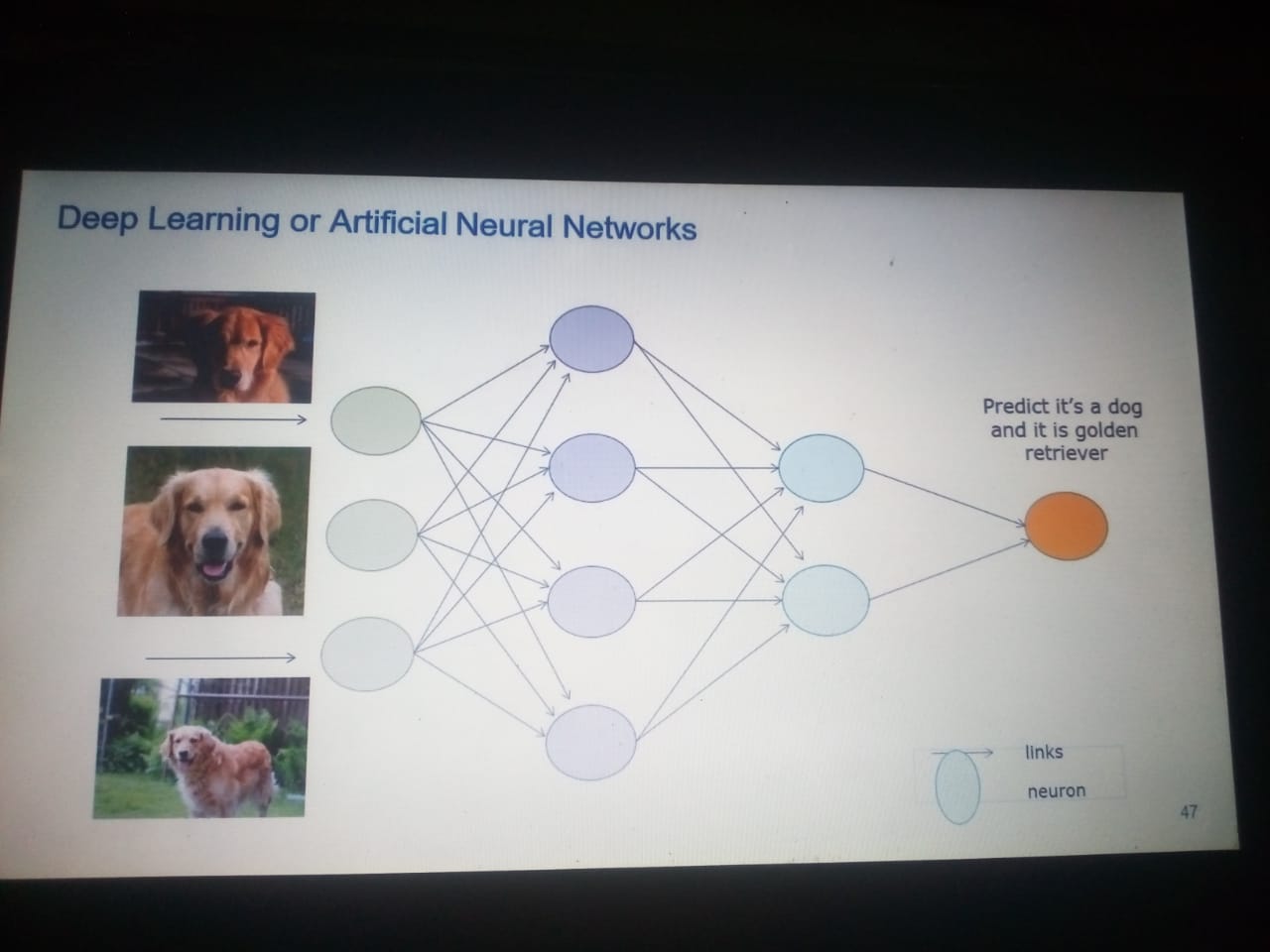
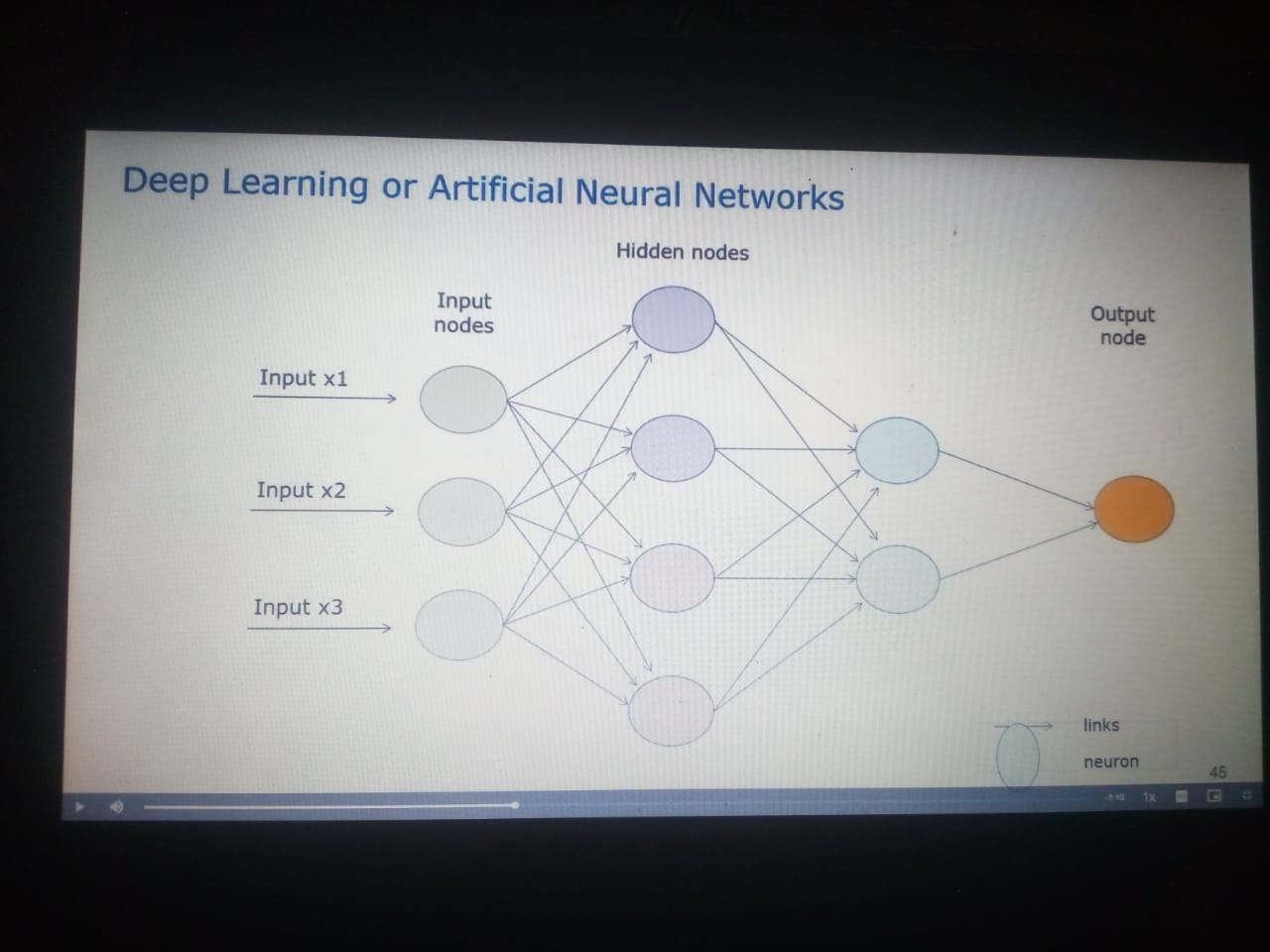
Image Recognition

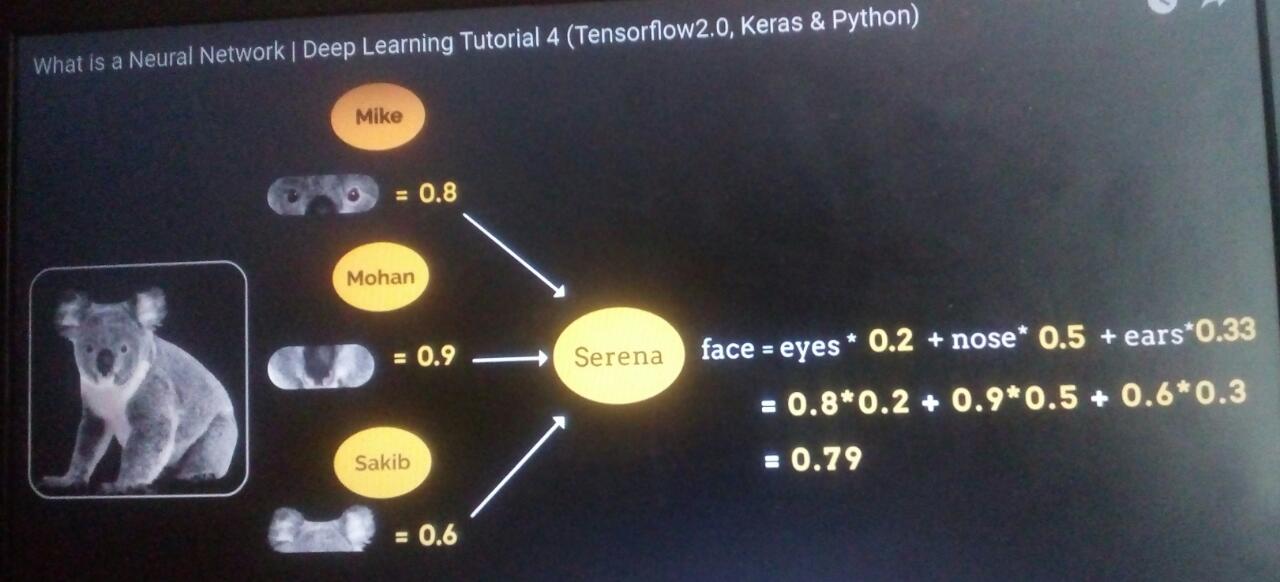
Computer Vision

Recurrent Neural Network(RNN)

Speech Processing

Text Processing





**DAY 8**

**Forecasting Analytics**

* Time Series Concepts
* Visualizations
* Transformations
* Moving Average Methods
* Exponential Smoothing Methods
* Model Building Methodology using ARIMA

**Resampling and Interpolation**

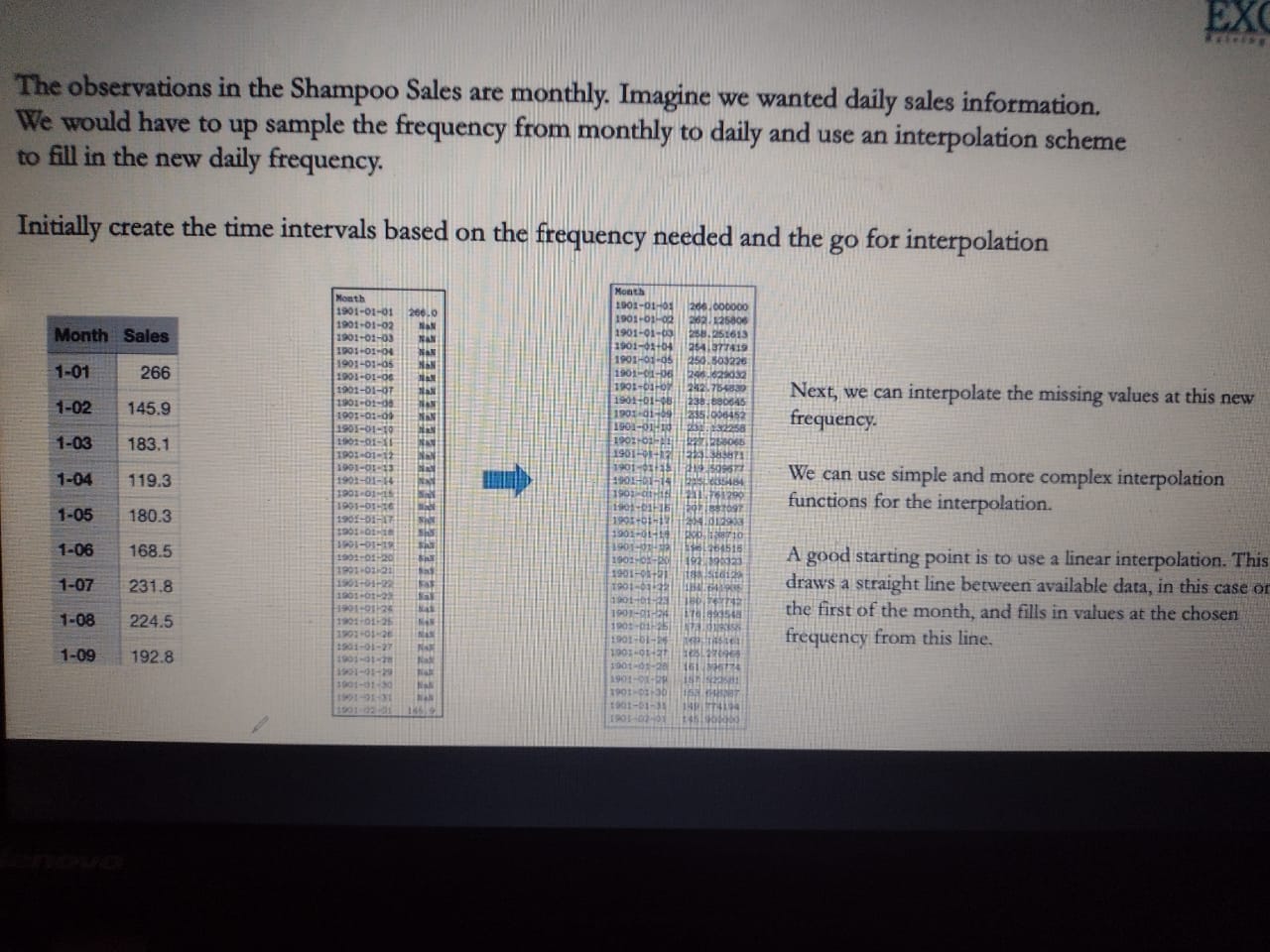
**Resampling: UP/DOWN**

EG. We have yearly data but my manager wants quarterly data its down sampling

Decades ko years me convert Up sampling

Decades ko Century me convert krna Down Sampling

**Interpolation:**



**Components**

**1.Trend**: In this component sales is increases Upward or downward or stationary

**2.Seasonality:** We can say this component is work like season wise If Christmas is there then chocolates sales will be high then again down then again in the Christmas season sales will be high

**3.Irregularity:** It does not follows any patterns like Up or Down

**4.Cyclic:** Repeating up and down ,don’t have fix pattern ,much header to redact

Test To check Stationarity

1.Rolling Statistics: plot moving average or moving variance

2.ADCF Test:Test Statistic and some Critical Values

Optimization Basics

Best solution, Like for the problem which will be the best solution this is known as Optimization techniques. Best of the problem you have to find out objective function that is this problem is maxima or minima

Two types of objectives

1.Maxima

2.Minima

Best on the problem solution we have to choose that

Gradient Descent Algorithm

Batch Gradient Descent: Use all m examples in each

Stochastic gradient descent: Use 1 example in each iteration

Mini batch gradient descent: Use b examples in each iteration

**DAY 9**

**Decision Tree**

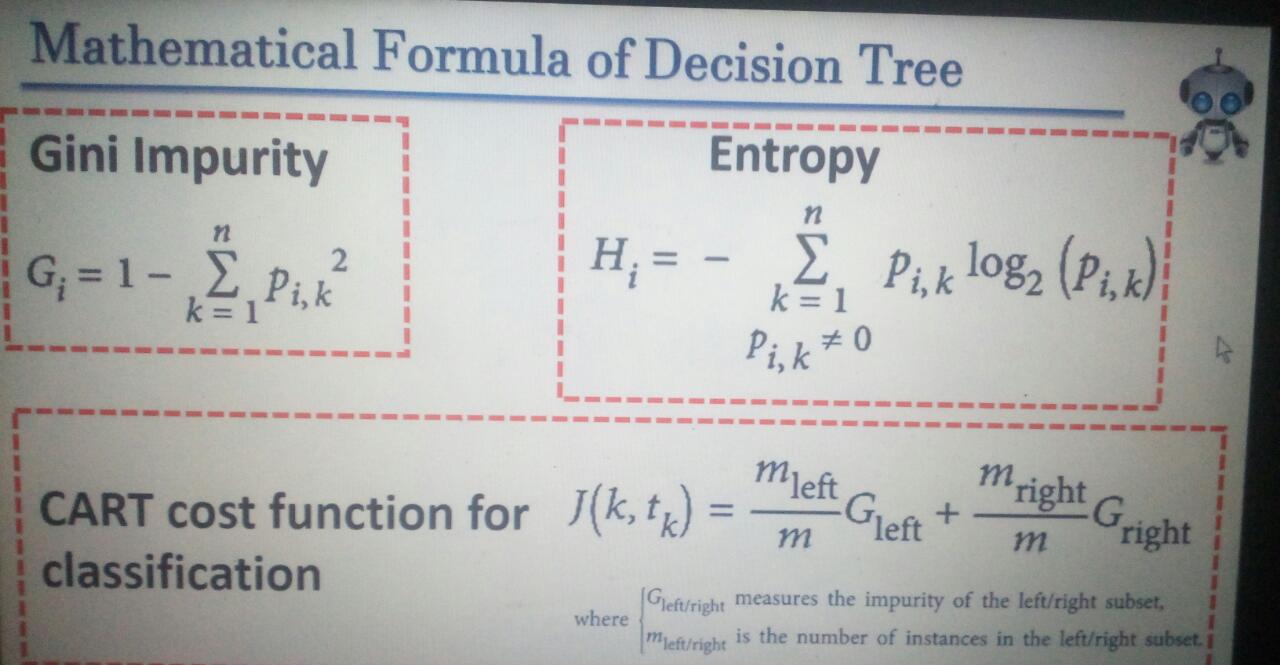
It is a type of supervised learning algorithm that is mostly used for classification problems and works for both categorical and continuous input and output variables

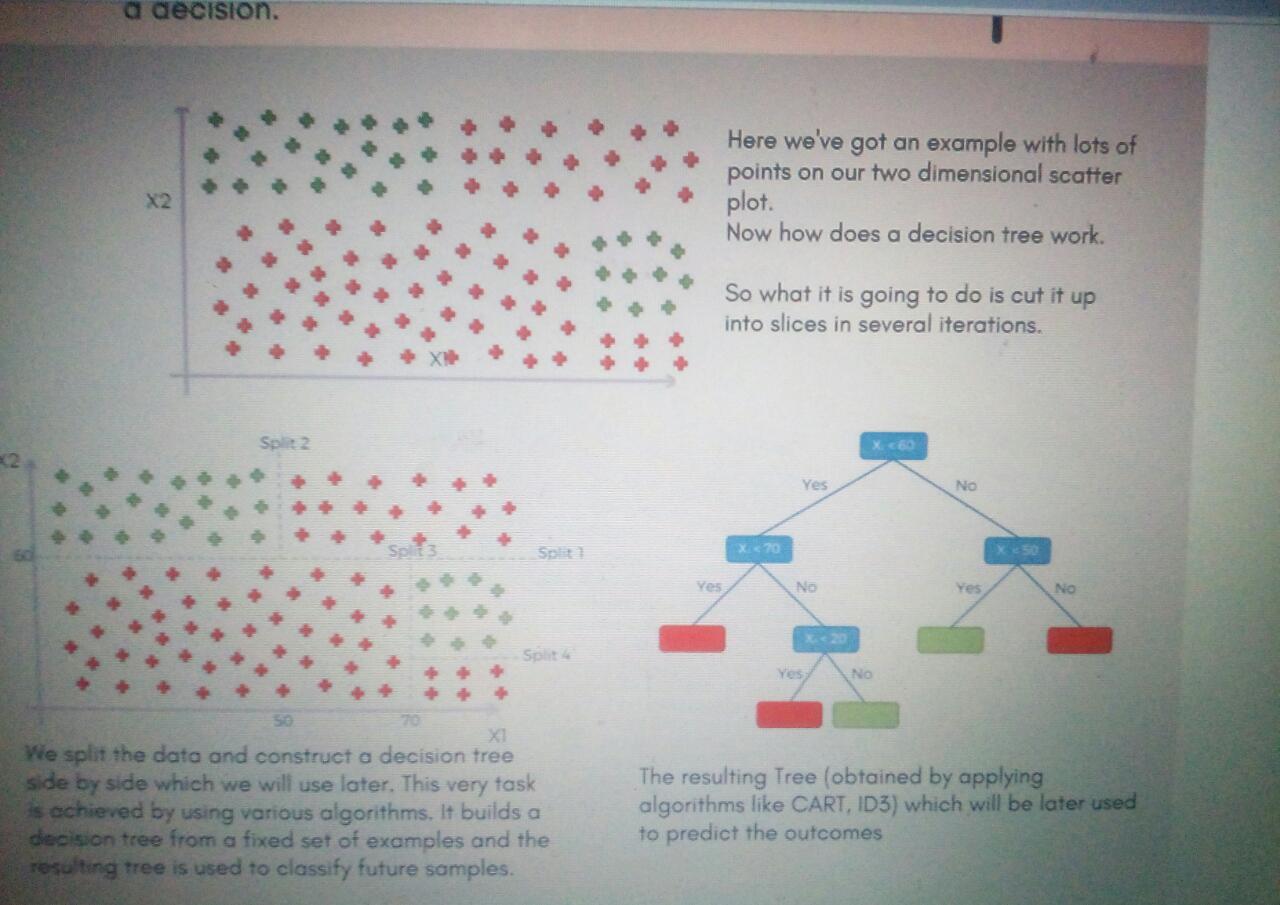
Using scikit learn uses the classification and regression tree ( CART algorithm to train decision trees

IDE tree also use for decision tree

A decision tree is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision.

* 1. Branch Node represents = Choice
  2. Leaf Node Represents = Decision





Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

**DAY 10**

**Random Forest**

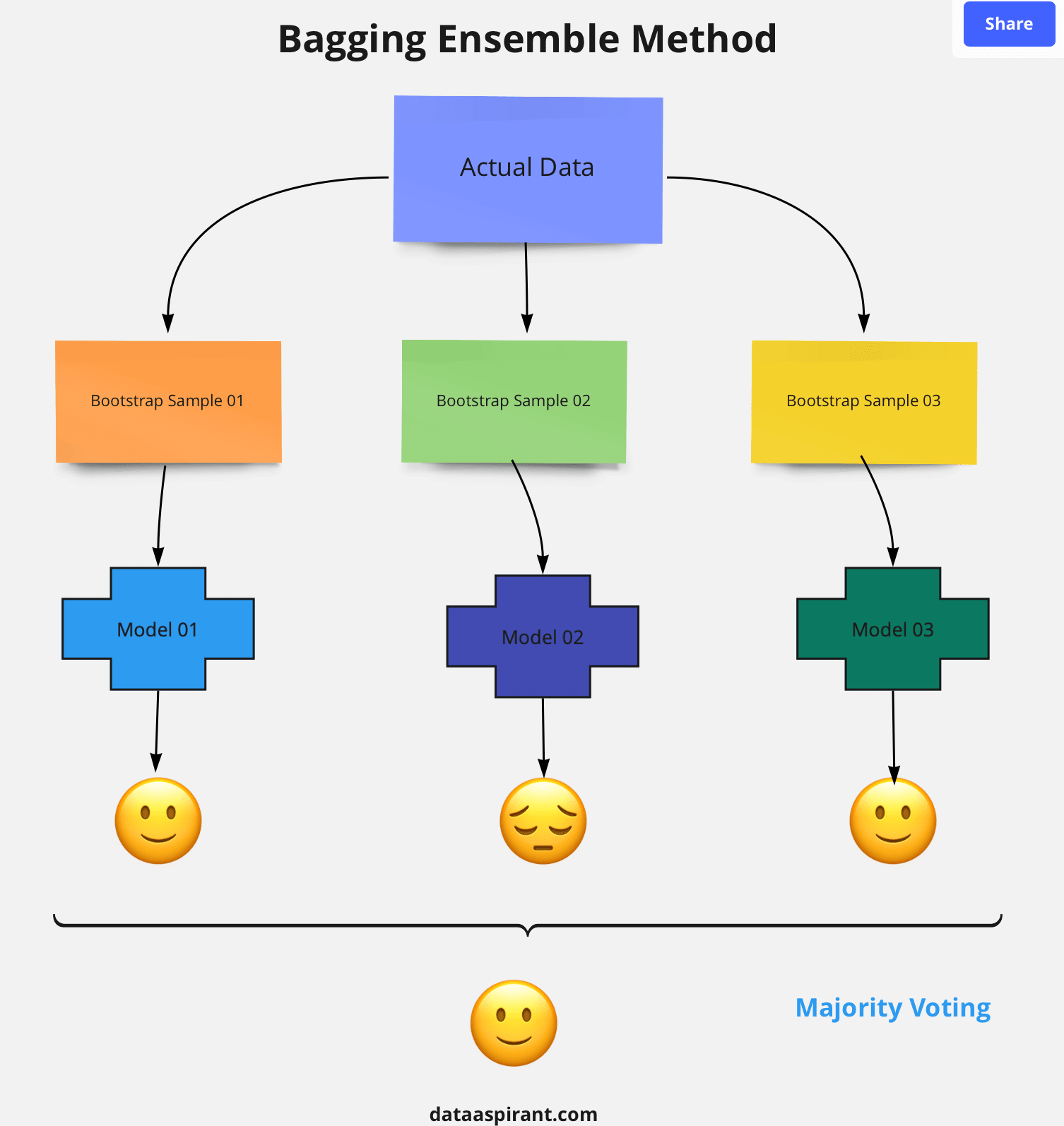
Before understanding the working of the random forest we must look into the ensemble technique. ***Ensemble***simplymeans combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

It is a type of supervised ensemble learning algorithm that is used for classification and regression problems

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction

*Ensemble uses two types of methods*:

* 1. **Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example,  Random Forest.



2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example,  ADA BOOST, XG BOOST

Prediction

1.Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stress the predicted outcome(target)

2.Clculate the votes for each predicted target

3.onsider the high voted predicted target as the final prediction from the random forest algorithm



**Use Cases**

This algorithm is widely used in E-commerce, banking, medicine, the stock market, etc.

For example: In the Banking industry it can be used to find which customer will default on the loan.

**Advantages and Disadvantages of Random Forest Algorithm**

**Advantages**

1.  It can be used in classification and regression problems.

2. It solves the problem of overfitting as output is based on majority voting or averaging.

3. It performs well even if the data contains null/missing values.

4. Each decision tree created is independent of the other thus it shows the property of parallelization.

5. It is highly stable as the average answers given by a large number of trees are taken.

6. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.

7. We don’t have to segregate data into train and test as there will always be 30% of the data which is not seen by the decision tree made out of bootstrap.

8.  It can handle binary, continuous, and categorical data.

**Disadvantages**

1. Random forest is highly complex when compared to decision trees where decisions can be made by following the path of the tree.

2. Training time is more compared to other models due to its complexity. Whenever it has to make a prediction each decision tree has to generate output for the given input data.

**DAY 11**

**K Means Clustering**

1.starting with Unsupervised

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. Unsupervised algorithms find patterns based only on input data. This technique is useful when we’re not quite sure what to look for

2.clustering Algorithms

Clustering Algorithms to do the task of dividing the populations or data into a variety of groups such that data points within the same cluster are similar to other data points within the same cluster that those in other groups. Basically, the aim is to separate groups with similar traits and assign them into clusters

3.K Means Clustering

Distance measured( Like Euclidean distance ) are used to calculate similarity and dissimilarity between the data points.Each cluste has a centroid.Centroid can be thought as the point that is most representative of the cluster.



**K-means Clustering – Example 1:**

A pizza chain wants to open its delivery centres across a city. What do you think would be the possible challenges?

* They need to analyse the areas from where the pizza is being ordered frequently.
* They need to understand as to how many pizza stores has to be opened to cover delivery in the area.
* They need to figure out the locations for the pizza stores within all these areas in order to keep the distance between the store and delivery points minimum.

Resolving these challenges includes a lot of analysis and mathematics. We would now learn about how clustering can provide a meaningful and easy method of sorting out such real life challenges. Before that let’s see what clustering is.

**Similarly, for opening Hospital Care Wards:**

K-means Clustering will group these locations of maximum prone areas into clusters and define a cluster center for each cluster, which will be the locations where the Emergency Units will open. These Clusters centers are the centroids of each cluster and are at a minimum distance from all the points of a particular cluster, henceforth, the Emergency Units will be at minimum distance from all the accident prone areas within a cluster.

**DAY 12**

**Hierarchical Clustering**

Pandas is use to manipulation the data like add delete clean

Groupby () is used on categorical data

df.groupby([“gear”])[“mpg”].describe()

**K fold** : a technique called Kfold cross validation which splits randomly the data into train and test set, trains the model using the train set and validates it with the test set, it will repeat this K times hence the name Kfold and takes the average error.

Recommendation System

Popularity bases : Based on formula or what is popular what customers like most that will it recommend

Content base: I cust like broduct A may br like product B and C

Collaborative filtering : Based on rating

Hybrid Recommendation syatem:

Mixture of all above time

Deep Learning

If the data is very huge so machine learning cannot handle huge amount o data in that condition deep learning plays important role. We don’t need to tells to model about prediction because deep learning has inbuilt algorithms which gives best prediction. It thinks like human brains

Why we need deep learning?

* Huge amount of data
* Complex Problems
* Features Extraction

What ?

* Handle Huge amount of structured and unstructured data
* Complex operation
* Problem Solved

Where?

* Medical Field: Nucleus prediction
* Robotics
* Self-driving cars: it will predict traffic signals, buildings
* Translation

ANN

Advantages: They store information on the entire network

They have the ability to work with incomplete knowledge

They offer distributed memory

Disadvantages:

The have huge hardware dependency

They sometimes have unexplained behaviour which can leave us tormented with results

There is no specific rule for determining the structure of artificial neural networks and appropriate network structure is achived through experience and trial and error

CNN:

Advantages:

They offer very high accuracy in image recognition problems

They are capable of automatically detecting important features without any human supervision

Weight sharing

Disadvantages

CNN do not encode the postion and orientation of object

They lack the ability to be spatially invariant to the input data

A lo of training data is required in order fot it to work efficiently

RNN:

RNN remembers each and every information through time

A recurrent neural network are even used with convolutional layers to extend the effective pixel neighbourhood

Disadvantages:

They have gradient vanishing and exploding problems

Training an RNN is a very difficult task

It cannot process very long sequences if using tanh or relu as an activation fun

PCA :

3 reasons or dimensionality reduction visualization,Interportability, and time and space complexity,we can visualize the data upto 3 dimension only if we have more than 3 features then for us its difficult to visualize the data then we used PCA its generate the PCA which have max info input it so we don’t have loss of data

If we apply PCA then How many PCA will be generated?

Large no of set of correlated variable into small no of reprented variable if we have 10 features the it will create 10 PCA but with high values

Like 1 st PCA will have the highest no of data

What is Random Forest and its work

What do you mean by ensemble machine learning techniques?

To combine multiple model for better accuracy

Random forest is an ensemble machine learning tech that avg is several decision tree on dof training set and aim of this to overcome the overfitting problem

What is bagging:

In this we will do sampling with replacement

What do you mean by gradient descent

Gradient mean s upward/ downward ,Decreasing something, it is a measure of property that how much output has change with respect to change in the input , minimization algo means decreasing something,we can used this boosting algo

Used in CNN ,minimize difference in the error.it will check that weather we want to upword and downword our model

Difference between correlation and covariance

Covariance is effect on one variable to another variable

If x is change then y is also change

Correlation How closely they are related with each other and what is the change if 2 variable have covariance

What is mean by ROC?

On x axis you have false positive rate and on y axis you have true positive rate it will give you 1 curve ,it compares various model so you can check how area under the curve so we can say that area is under the curve is better to predict the result

What is mean by regularization and why its needed?

Used to avoid overfitting ,in this we add extra penalty term and its improve the extra term that is penaulty term its used for regularization

Regularization

* 1. Lasso
  2. Ridge

Hyper tunning

Gradient descent

This technique use for handling overfitting and underfitting the data in regression

Bagging

Boosting

Is used for classification

Feature Engineering

Univariate Selection

RFE ( Reverse Feature Elimination)

Tree based feature Selection

Selecting those feature which is more contribute for the prediction

Objectives :

1. Reduce Overfitting
2. Improve accuracy
3. Reduce training time

Mean Absolute Deviation