Learning Analytics Tools

Professor Ramkumar Rajendran

Educational Technology

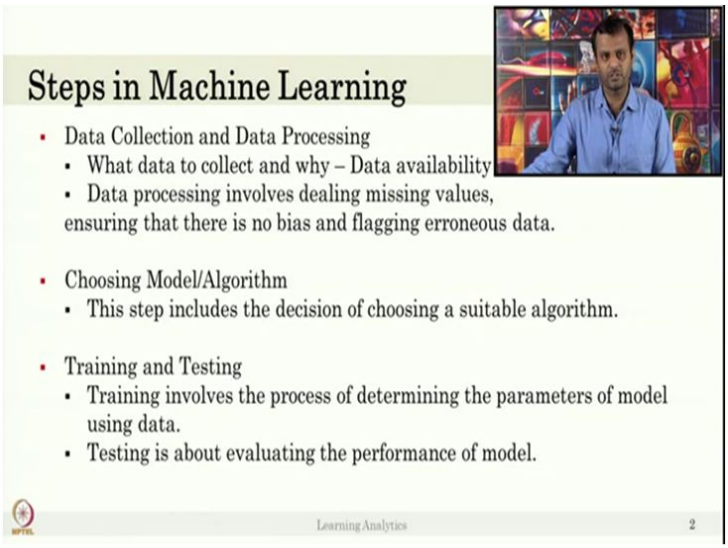
Indian Institute of Technology, Bombay

Lecture 3.2

Introduction to Machine Learning Part 2

Welcome to the Learning Analytics Tools course. This is the introduction to machine learning part 2. So, in machine learning there are 3 steps.

(Refer Slide Time: 00:30)



Step 1

The first step is data collection and data processing. This is similar to what we saw in learning analytics. In this step, we have to decide - what data to collect and why we have to collect and importantly we have to make sure that the data is available or not. Also, you have to do the data pre-processing like removing the missing values or outliers. Ensure that there is no bias in the data. Also if you find any errors you have to remove those data points.

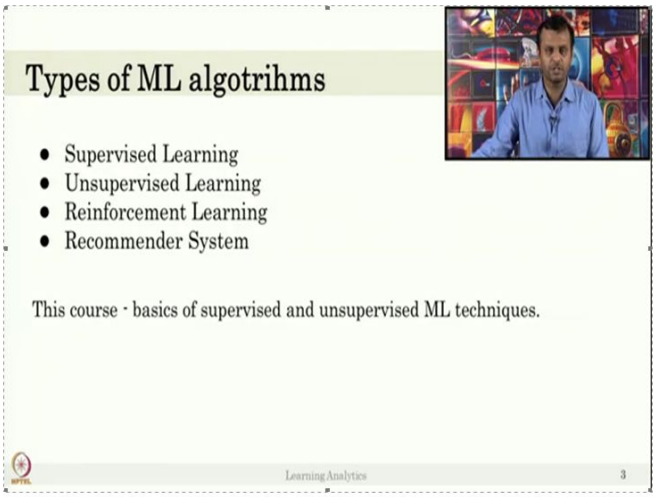
Step 2

The second step is choosing a model or algorithm. This is like finding which model to use for the data on the research question you are selected. This step includes the decision of choosing a suitable algorithm.

Step 3

The third step is training and testing. So now we have data. You collected the data. You have a research question and you also have the model to use it for your analysis and the next thing is you have to train and test. So you have to classify your data into training and testing data and use the data training and testing. Training involves the process of determining the parameters of the model using the data. Testing involves evaluating the performance of the model. You can create more than one model and you can compare the performance of the models using the test data.

(Refer Slide Time: 01:55)

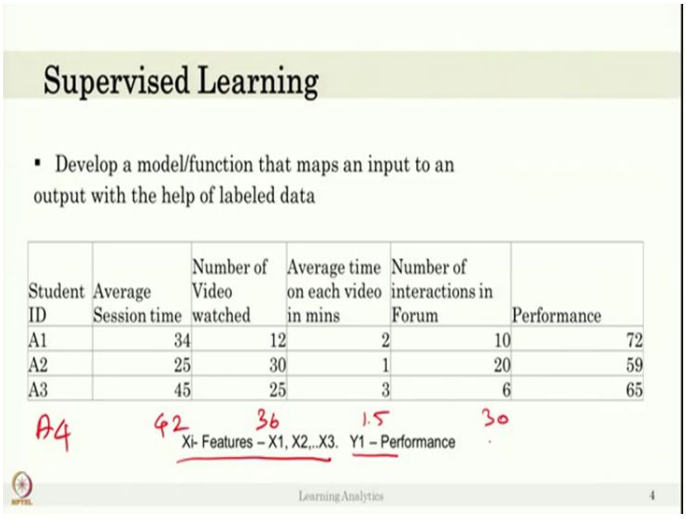


So let us look at the type of ML algorithms. There are 4 types: supervised learning, unsupervised learning, reinforcement learning and recommender system. In this course we will talk about supervised and unsupervised machine learning techniques.

So what is reinforcement learning? Reinforcement learning mostly concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

Recommend system as the system which in the education context make recommendations based on the student's interaction. It recommends what to do next like provides you with the questionnaire or new content or new example something like that. Recommender system in another context, it can be recommending a new algorithm itself or recommending a new program or new strategy something like that.

(Refer Slide Time: 02:38)



Let us start with the supervised algorithm. The supervised algorithm is developing a model or function that maps an input to an output with the help of labelled data. So you have to find the mapping between the input and output. What is input and what is the output here, let us see.

For example, you have collected the data of students, 3 students here. You got a data of 3 students and your data that is average session time. Imagine they are working on MOOC or a Mettle kind of TELE. Let us say it is a MOOC. First, you have to see the average session time. So in a MOOC, they might be logging into the course several times.

What is the average session time for the student A1 when he is interacting with MOOC?

It is say 34 minutes. It is given in minutes and the number of the videos the student watched overall maybe 12 or 30 or 25 that is a number you want as the second feature.

The other input we can have is the average time on each video in minutes. Like you watched 12 videos but what is the average time you spent on each video? Say for 2 minutes. So some videos you might have watched more than 2 minutes, some videos were watched less than 2 minutes but the average is computed.

Also, we can talk about the number of interactions in the forum like a 10, 20 or 6. So I assume that there are 4 features I can collect from the MOOC data.

I have a log data, I computed these 4 features using excel or written some script to capture it from the raw data. Also, I have the students performance in the final exam or the exam conducted after the MOOC course. So I have the final exam here.

I will consider this input, which I can observe like

$$x_1, x_2, x_3, x_4$$

or as a feature like

$$x_i$$

The i here is 1, 2, 3, 4 features and I will consider this performance as a predictive value or a dependent value. This performance depends on these 4 values that is why it is called a dependent variable. These are all independent variables.

So we call that as a y that is the performance as the label y and x_i denotes the features x_1, x_2, x_3, x_4. Here I have students A1, A2, A3. Now we want to create a model for this data so that if a new student A4 comes with say a different time 42 and number of videos watched let us say 36 and you watched for only 1.5 minutes, number of interactions say 30. What will be the performance?

$$x_1 = \mathbf{42},$$

$$x_2 = \mathbf{36},$$

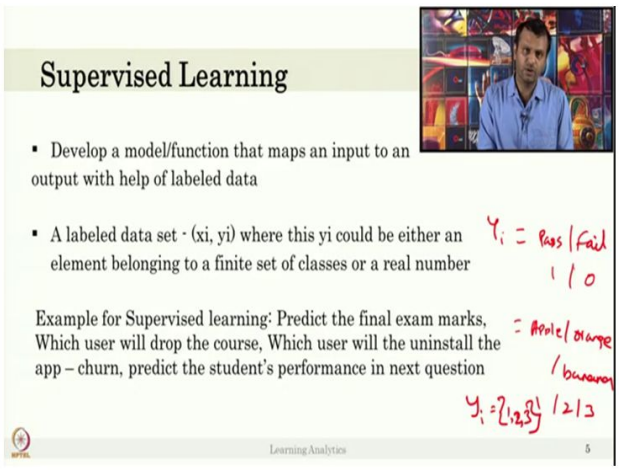
$$x_3 = \mathbf{1.5},$$

x_4 **=30**

So the idea in machine learning is that you have supervised data that is you have students interaction that is observation input data. Also, the labels to create a model using this data and test it or predict the performance on the new student or new data we are yet to see. That is what, supervised machine learning means. Now let us understand what is x, y,

x_1 **,** y_1 means.

(Refer Slide Time: 05:54)



So a labelled data set has duplet (x_i **,** y_i **)**,

where this y_i could be either an element belonging to a finite set of class or a real number.

x_i as seen in a previous slide are 4 features we collected from the MOOC data interactions, y_i over there is performance or student score.

The y_i can be pass or fail or the real number as we saw in the last example. Pass can be denoted as 1, fail can be denoted as 0 or it can be apple, orange or other fruit. We can denote this as 1, 2, 3. So that now y_i is set of 1, 2 and 3 in this example. Similarly, y_i can be 0 or 1(i.e. binary). It can be a real set of numbers or the natural numbers as we saw in the previous class like performance in the exam at 63, 40 or 70 or something like that.

So examples for the supervised algorithm is that we can predict students final scores based on the interactions in the class or in the learning environment.

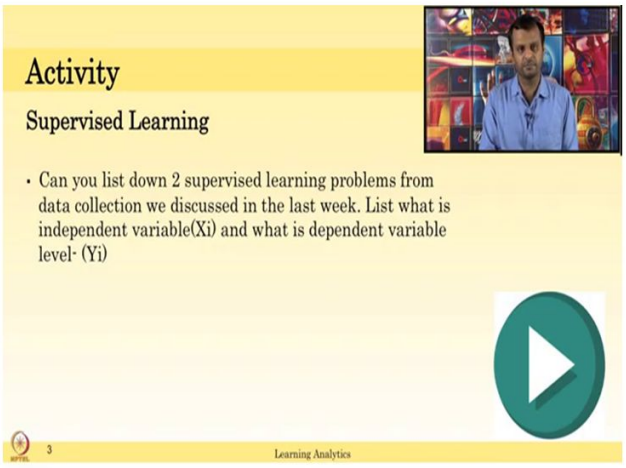
We can predict in MOOC which users will drop out, drop the course in 5 weeks or 6 weeks something like that. The dropping will be binary classification saying that you will drop or not drop or which user will uninstall the app that is churn rate. Suppose considered you are creating the education mobile app and we want to know which user will install the app and when they will uninstall. Can you predict it based on their interaction with the app? So then you are predicting whether uninstall\install churn rate(that can be again 0 to 1 binary classification problem).

Predicting the student's performance in next question that is a more fine level prediction that is based on the student's interaction with the system for say last 10 minutes or last couple of sessions and you want to predict whether the student will answer the next set of questions given to him/her?

If the student is not going to answer you might consider giving the less difficult questions or coding hints or asking them to read something that will help to save the student's time and also the student can learn better. So these are examples of supervised learning. In these all examples, we know we want to predict the student's performance something like a final exam score. We should have the final exam score. We can collect it.

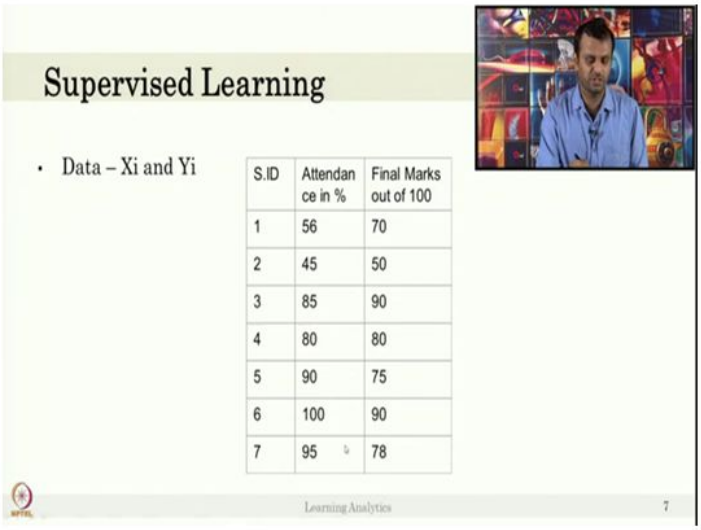
Also, we want to know other students dropped the course or not. You know what to predict. You know what is the value you are going to predict, also you have to come up with the set of input features like x_1, x_2, x_3, x_4. So supervised learning you know what is the input also what to predict. If you have both x_i and y_i then it is considered to be a supervised learning approach.

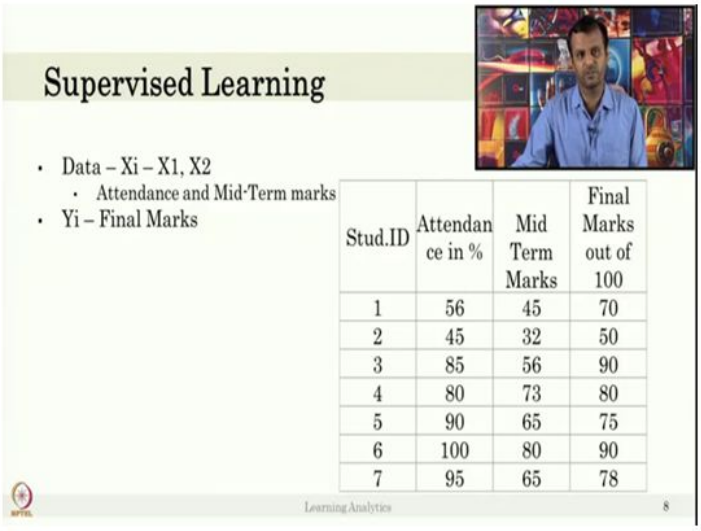
(Refer Slide Time: 09:00)



So given the examples, you saw in the previous slide can you list down other two examples for supervised learning problems from the data collection we discussed in the last week. Just what is the independent variable? That is x_i. What is the dependent variable? That is y_i, that is the label. List down 2 problems and list down its independent variables and dependent x_i and y_i. After listing it down resume it to continue.

(Refer Slide Time: 09:30)



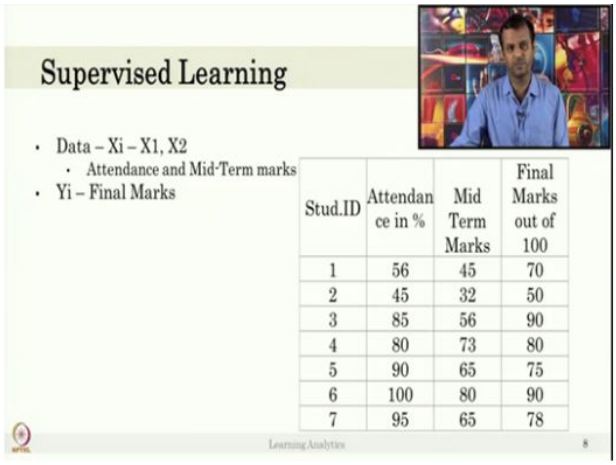


So you might have listed down some of the examples like x_i and y_i. So let us consider that you have x_i and y_i. Say there is a simple example students ID there are say 7 students and we have the attendance percentage in the semester. Also, we have their final marks in the semester.

Now, this is x_i and y_i.

x_i = **attendance percentage**

y_i =**final marks**



I want to predict the student's final map score final marks in the exam using the attendance percentage so or you can have a multiple-input value say x_1 and x_2. For example, attendance and mid-sem marks. So I have 2 more, 2 data input features. I want to predict the student's final marks. So that is y_i . The y_i is the final marks, it is a real number.

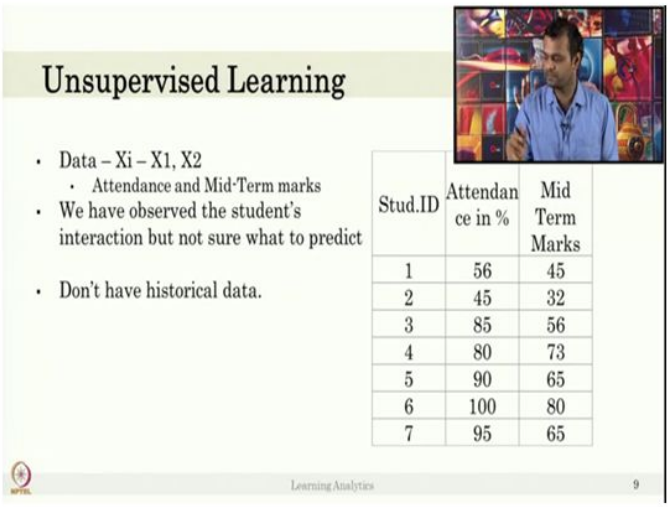
x_1 **= attendance percentage**

x_2 **= mid Term marks**

y_i **=final marks**

So this is a supervised learning technique and should be used for the prediction.

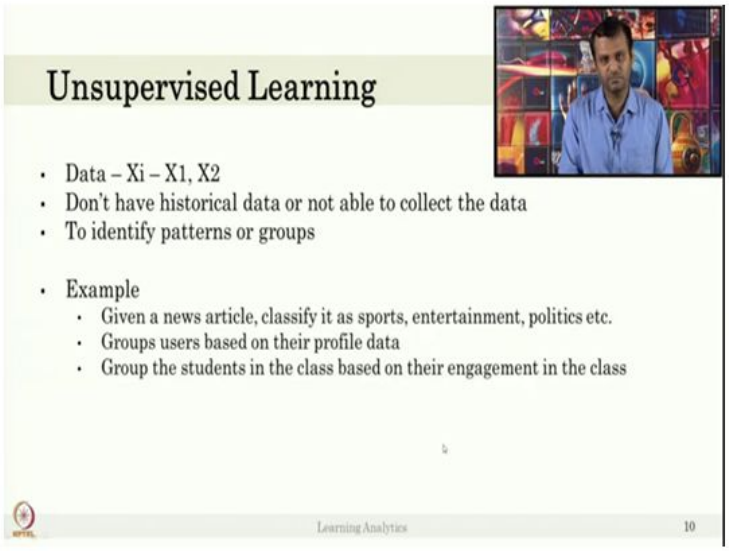
(Refer Slide Time: 10:34)



So what is unsupervised learning? So we said that supervised learning has both x_i and y_i. Unsupervised learning has only x_i. Is like we have observed the students interaction but not sure what to predict or we do not have historical data.

That is that in the previous slide we saw that students attendance, mid-term marks and exam scores are available but that data is available from your previous years teaching, record or data available as some students have already taken the exam. You want to predict the student's performance in the current batch. You do not have any historical data of the score or something. In that case, we can use the unsupervised learning approach. So in unsupervised learning approach, we have x_1 and x_2. There is no y_i. We do not know what we are predicting but we have collected a lot of data. We want to see is there any pattern evolved from the data or any clustering happening in the data.

(Refer Slide Time: 11:35)



For example, given a news article, classify the news article as sports, entertainment and politics. In the Google news, it happens automatically without having labelled it as, or classifying a sports or entertainment.

But if we have a lot of human data labelled, then this is the supervised problem.

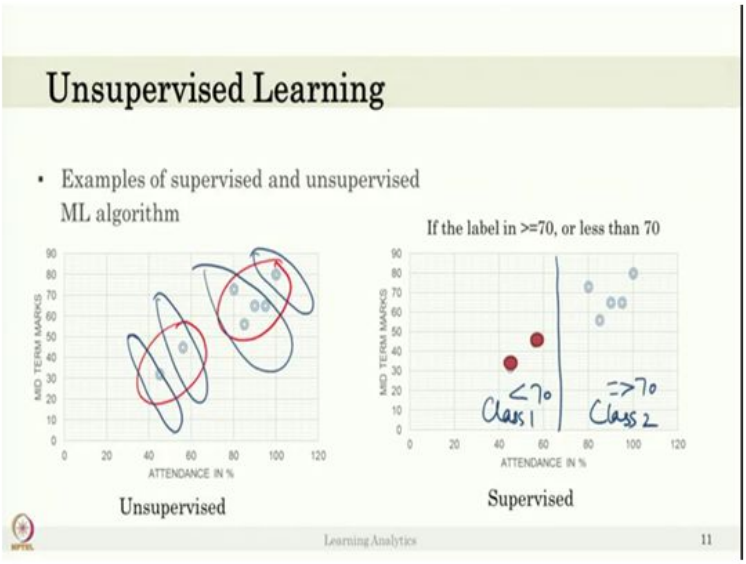
Consider there is a given news article. A topic of one news item says there was a festival in a particular state and we want to collect all the related article from different newspapers. If you want to collect about that particular news from a different newspaper, news publisher, then it can be an unsupervised algorithm because we may not know the label of this ahead of the time. So there the system automatically groups similar articles. For example, a festival in the state Uttarakhand and if you want to consider that news it uses the keywords in those news articles and tries to find is there a similar keyword in any other news magazines or newspapers then it collects everything together and put it under one particular news heading.

If you go to Google News and if you use news.google.com you will see this kind of grouping together happens using unsupervised learning algorithms and other one is group users based on their profile data.

Suppose you want to create a project lab course or something like that and you want to group the students based on TOEFL data like based on the previous background or the branch in their class 12 or diploma or some other thing or the background information, you can group that

information and you can create groups and assign some tasks to them or group the students in a class based on engagement in the class. There are some students who are highly engaged, some students who are not engaged you might find a grouping of these students and you may want to mix them to make peer learning better or something like that.

(Refer Slide Time: 13:50)



So in a nutshell the difference between unsupervised learning and supervised learning is, consider we have attendance in percentage, midterm marks. If I plot that in X and Y axis and I have plotted like this and I have no idea where a particular student belongs to. So if I apply a grouping algorithm or clustering algorithm I can group this as one group and this as a second group or I think if I want to apply further these 2 groups may not be right. If I want to apply 3 groups maybe I should go and create new groups. Maybe this is the one group. This can be the second group and this can be the third group and this can be the fourth group.

How many clusters are good, 2 or 4? That we have to identify by using the mechanism of errors or minimal errors. We will talk about that later in a clustering class. So given this data without any label, we can group them into 2 groups or 4 groups. So clustering algorithm can identify the groups with similar behaviour in the data.

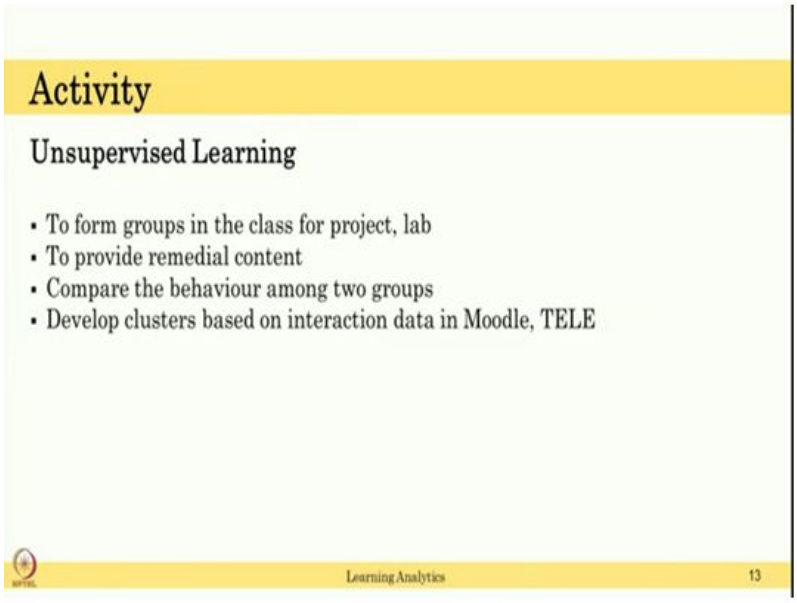
Whereas in the supervised algorithm we know that there are 2 groups here. These students have scored less than 70 marks in the final exam and these students have scored more than 70 marks in the final exam. So this is a exact 2 class classification problem. This is a class 1. This is class 2, class 1 is less than 70 marks and this is equal to or greater than 70 marks.

So now you know the label y and you know the x_1 attendance and x_2 a midterm marks. Now, this is a supervised learning problem. In this problem, we have x_1 and x_2 we do not have y. So it is not a supervised problem. We can come up with the clustering. Say 2 clusters or more than 2 clusters depending upon the data and the interaction behaviour how they interact with each other. We will talk about the clustering in details in a separate class. So this is to introduce the difference between supervised and unsupervised learning.

(Refer Slide Time: 16:03)

So since you have seen supervised and unsupervised learning, can you list down 2 unsupervised learning problems from a data selection we did in the last week. After listing down please resume the video to continue.

(Refer Slide Time: 16:19)



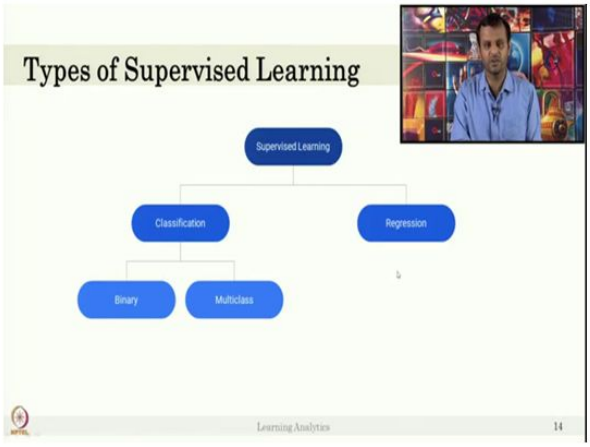
So as I mentioned already you can use that unsupervised algorithm to form groups for a class project or form lab projects. You may not want to form a group with a similar set of students instead you want to mix and match.

For that you first you need to find what are the student's behaviour or you can use this unsupervised algorithm to provide a remedial content or extra coaching to them or you might want to give exam which is less difficult or teach a special course or something like that or you can compare the behaviour among 2 groups.

If you have a group A and a group B and you can identify patterns among these 2 groups and compare the patterns of group A and group B. The patterns can be like order actions they do. You can identify patterns in the unsupervised algorithms. Also, you can develop clusters based on interaction data in Moodle or TELE.

So the interaction data we can use it to create a clustering algorithm and create clusters to group the students into multiple groups, and you can provide a different level of recommendations to them.

(Refer Slide Time: 17:27)

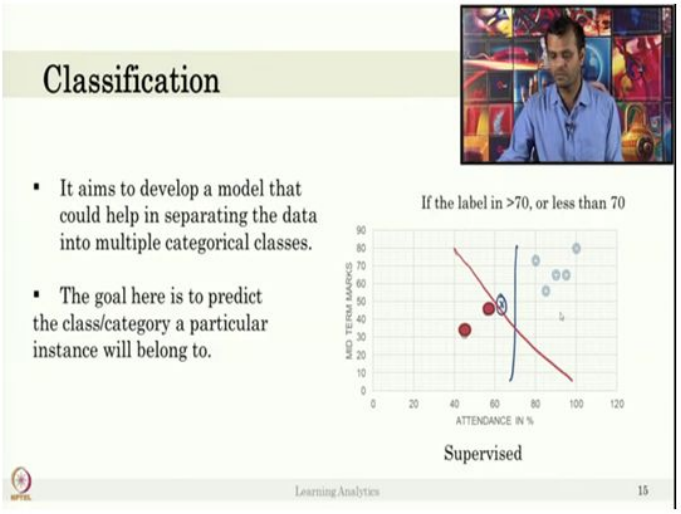


So here we have the types of supervised learning

1. classification
2. Regression

Unsupervised learning classification again goes to binary classification and multiclass classification and we saw that binary classification has 0 and 1. Multiclass can a 1, 2, 3, 4 or apple, banana or something like that. In regression, y is usually the real number like performance 65, 70, 75 something like that.

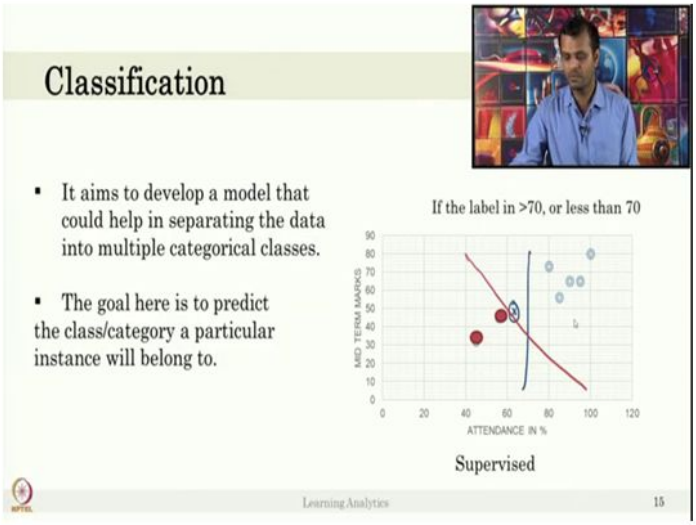
(Refer Slide Time: 17:52)



So in a classification problem, it aims to develop a model that could help in separating data into multiple categorical class. In the case of two classes, it is a binary classification problem. The goal here is to predict the class or category to which a particular instance will belong to.

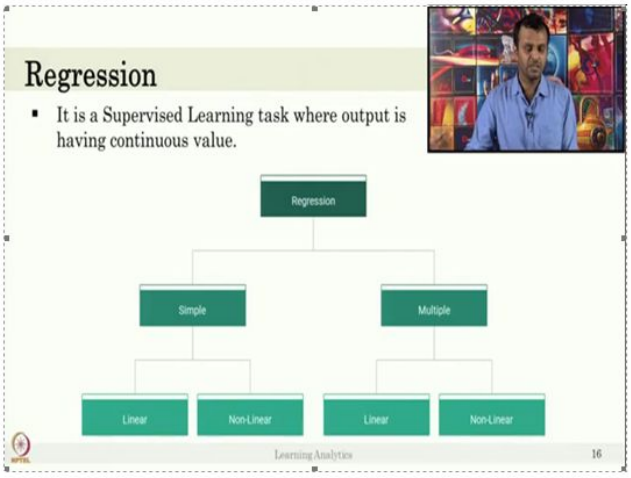
For example, if I have new data, say new data with attendance 63 and the midterm marks is, say 40 or 50, where this data belongs to? Do these data belong to class 1 or class 2? It is based on where you draw this line, right. How you draw the line? We will talk about that in detail. Suppose if I have a line drawn like this, this data might belong to the other class. So it depends on the where you draw the line will talk about that in detail but given a classification algorithm when a new data comes, the goal is to put the data into either of the class.

(Refer Slide Time: 19:06)



In regression, again it is subdivided into simple and multiple. In simple it is a linear, non-linear and also in multiple it is a linear, non-linear. We will talk about a simple linear regression now.

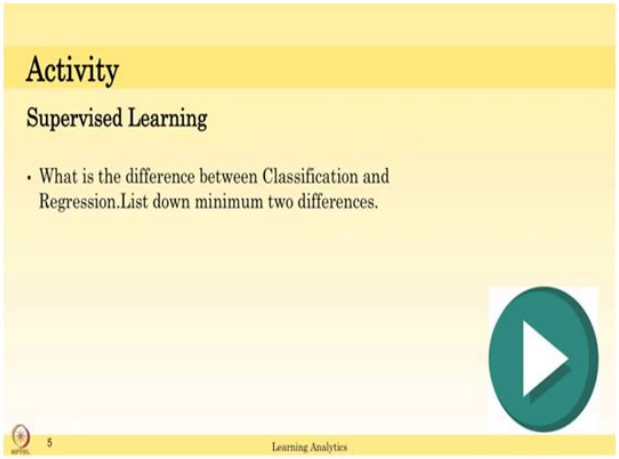
(Refer Slide Time: 19:19)



The output is the continuous variable. This is the same plot we discussed a few slides ago. There is attendance and the final marks. If I draw a line, a linear line which fits all these data which assumes that there is a linear relationship between attendance and final mark and I hope to fit these values in a linear line. So the simple linear equation aims to fit/map the x_i to y_i. So this is x_i and y_i. Let us try to fit the line between these 2. It assumes that there is a linear relation between these two variable. Consider there is a new data or new student with attendance equal to 65%. So what will be the students mark in the final exam?

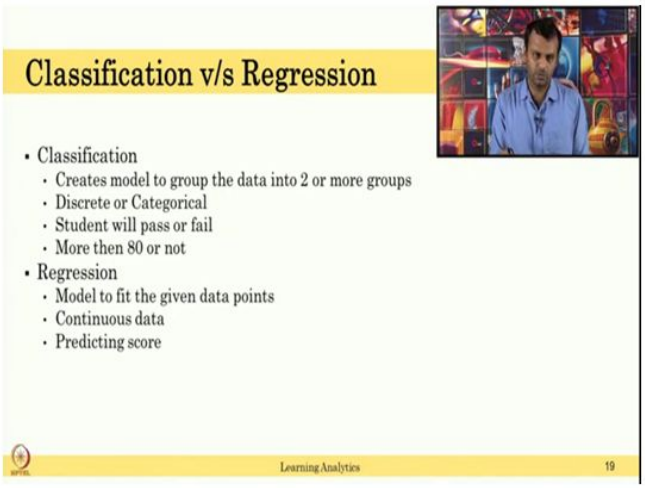
So since we fit this line using a linear regression algorithm, we will able to tell what will be the student score in the final exam. Since it is 65 we can draw the line and the mark will be around something like 68 or 69 or something like that. So in a linear regression algorithm instead of classifying in the new student's data into either the class 1 or class 2 or class 3, we are trying to find the score using the continuous variable.

(Refer Slide Time: 20:44)



So you have seen just the introduction of classification and regression. We will talk about this classification and regression algorithms in detail but given the introduction of classification and regression algorithms can you list down minimum 2 differences? After listing it down resume the video to continue.

(Refer Slide Time: 21:04)

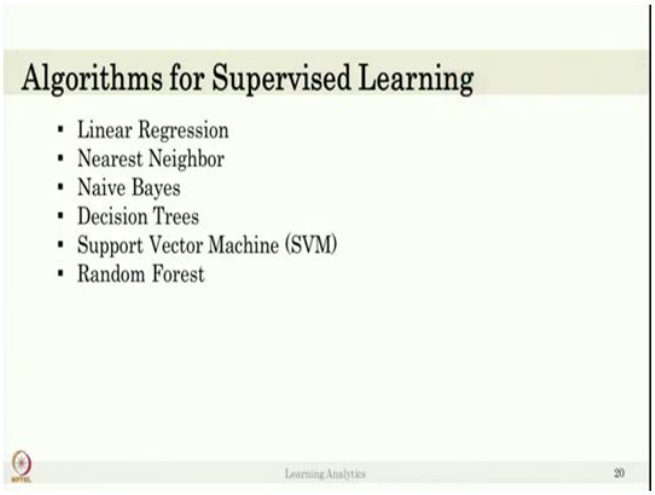


So the classification algorithm creates model to group the data into 2 or more classes. The data y is usually the discrete or categorical. Like example student will pass or fail, student will get more than 80 marks or not, something like that.

Whereas in regression the model trying to fit/map the given data points x_i to y_i. Also, y is continuous data its a real number and regression is about trying to predict the score. For example,

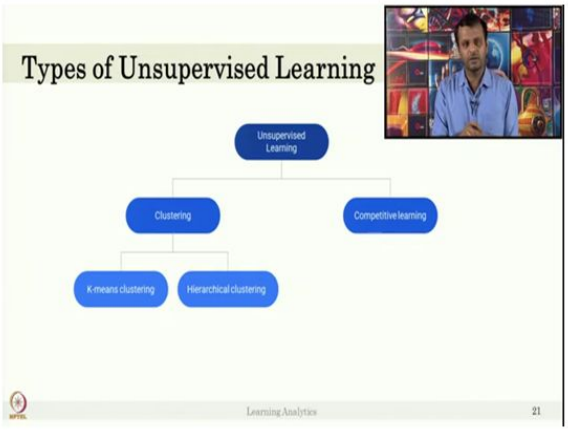
a new student comes what will be the score instead of trying to put him into whether the student will pass or fail or the student will get less than 70, more than 70 or more than 80, we are trying to predict the score. This is the difference between these 2.

(Refer Slide Time: 21:54)



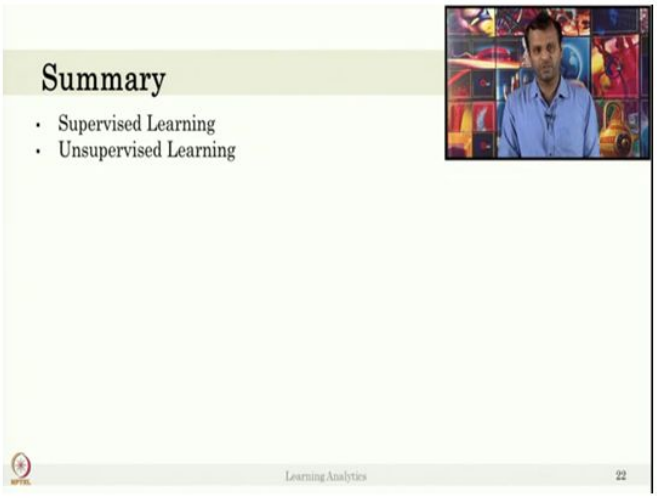
So there are few algorithms for supervised learning that is linear regression, nearest neighbor, Naive-Bayes, decision tree. We will see these algorithms in detail not support vector machine or random forest but we will see the first 4 algorithms in detail.

(Refer Slide Time: 22:07)



For unsupervised learning, there are 2 types again clustering and competitive learning. Like we saw in the supervised learning there is a classification and regression, in unsupervised learning, there are 2 types like clustering and competitive learning. In clustering again there are many types of clustering. We just gave 2 of them here. So K means clustering and hierarchical clustering. We will discuss these in detail.

(Refer Slide Time: 22:29)



To summarize, in this video we saw

1. what is supervised learning?
2. what is unsupervised learning?
3. what are the types of supervised learning algorithms?

In unsupervised learning we saw a clustering technique. We did not discuss much but we introduced what is clustering. We will discuss each of this algorithm with example algorithm in detail in further classes. Thank you.