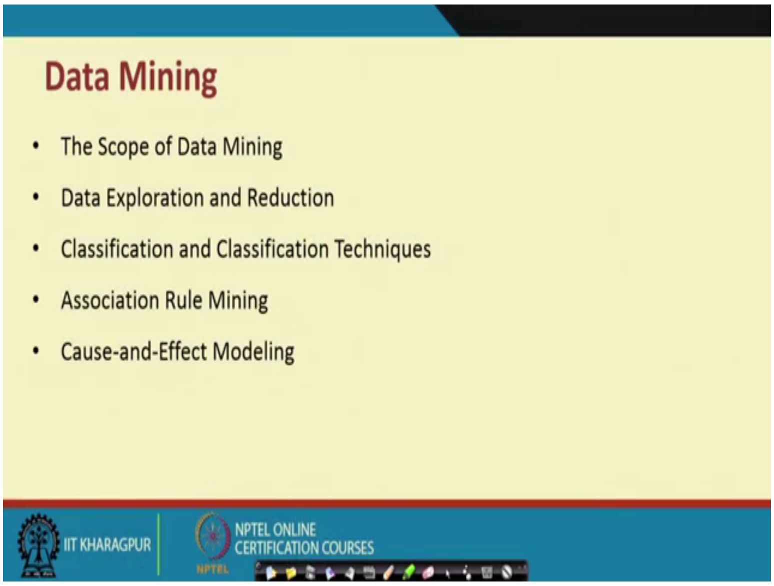
Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 39
Predictive Analytics: Data Mining

Hello everybody this is Rudra Pradhan here. Welcome to BMD lecture series, and today we will continue with the predictive analytics, and that too coverage on data mining. In fact, this particular technique or this particular tool is very vast, and very useful for business problems; where we have lots of complexity and the kind of dynamics.

And with the help of this particular technique again we like to understand the particular data structure, and then create a kind of environment through which we can do better predictions and better forecasting's. In fact, in the data mining we have a couple of items through which you can actually analyze the kind of problems, and then come out with a kind of solution through which the particular decision will be very effective.

(Refer Slide Time: 01:06)



Data Mining

- The Scope of Data Mining
- Data Exploration and Reduction
- Classification and Classification Techniques
- Association Rule Mining
- Cause-and-Effect Modeling

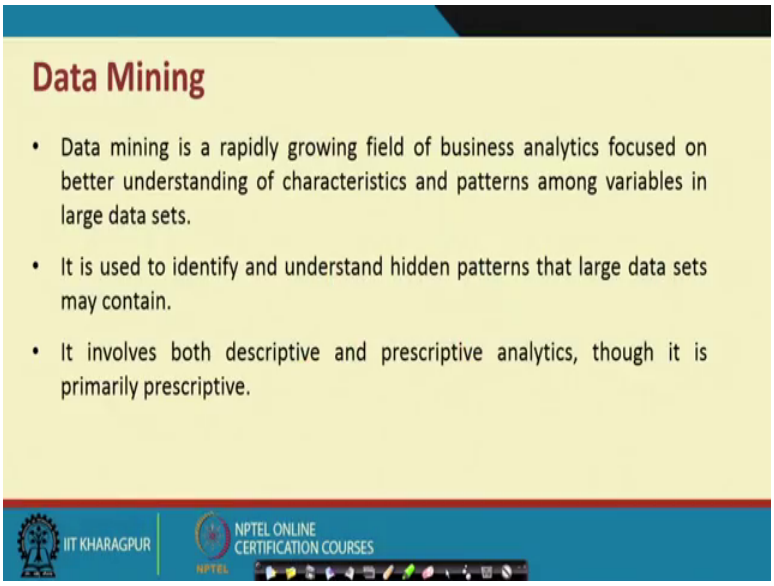
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

In the data mining so, the typical discussion will be the scope of data mining, then we will go for data explorations, and data reductions and again so, classification and the classification techniques. So, it is with respect to association rule mining and cause and effect modeling. In fact, we have already discussed couple of predictive analytics starting with a simple regressions that association, classification, cause and effect modeling in the context of you know regression modeling.

So, that means, altogether we have already discussed certain techniques in the predictive analytics, which are actually part of this particular data mining process. But here one thing we like to give high stress on the kind of cluster analysis, again this is a kind of classification.

So in fact, we have already discussed so many classification techniques like random forest and support vector machines, and here we have a kind of beautiful classification structure through cluster analysis, and for this kind of environment; means, business environment this particular classification technique can give better kind of inference through which you can understand the problem more accurately. And then we will create a kind of structure through which you do the prediction and the kind of forecasting's. So, typically we start with what is exactly the data mining.

(Refer Slide Time: 02:46)



Data Mining

- Data mining is a rapidly growing field of business analytics focused on better understanding of characteristics and patterns among variables in large data sets.
- It is used to identify and understand hidden patterns that large data sets may contain.
- It involves both descriptive and prescriptive analytics, though it is primarily prescriptive.

IIT KHARAGPUR
NPTEL ONLINE
CERTIFICATION COURSES

And then we will connect with some of the techniques through which you can actually understand the kind of data structure. And then analyze the problem as per the particular management requirement.

So, what I have mentioned already. So, it is a kind of beautiful technique through which we business analytics can be focused through which to understand the particular business problems as per the big data set, and the kind of complex number. If you know, means multi multiple number of variables through which you can analyze the particular problem. It is used to identify and understand the kind of hidden insights through in the data and the kind of the kind of variables which you have identified.

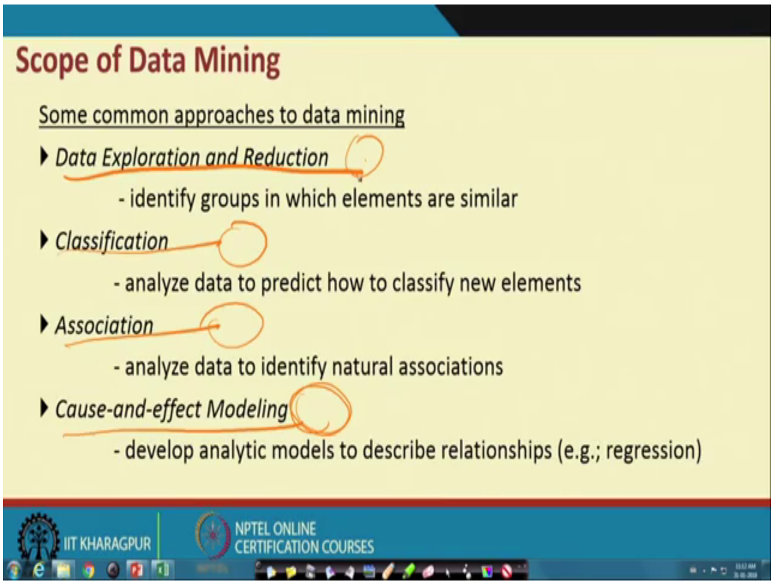
And then we develop a kind of set up and the structure through which you do the kind of prediction, it is a actually more or less same like neural network structure random forest and support vector machines or any kind of regression kind of techniques; that means, actually the fact is that we have a large data, and that too with a more number of variables.

And in one instance we have a kind of structure then using the data we like to just test and verify and find out the particular moulding set up through which you can better prediction and better forecasting for the future. And in another environment sometimes we do not find a any kind of insights in built insights initially. And through which you can just verify and then develop a kind of set up through which you do the prediction.

But in the second case typically the structure is it to develop a kind of system, and set up through which we have to again trained then verify then test and then finally, do the kind of prediction and forecasting as per the particular requirement. And cluster analysis and that too data mine in general data mining process is like that to understand the particular data set, and the understand the particular problem with respect to more number of variables. And the kind of dynamics, and then develop a kind of system and structure through which you can actually go for better predictions better forecasting's as per the particular management requirement or problem requirement.

So, it also involves both descriptive and prescriptive analytics; that means, the things which you are discussing right now is a kind of predictive analytics structure. But what I mentioned earlier that all analytics are very closely connected. And in fact, we are still we are discussing predictive analytics. But we need actually descriptive structure and predictive structure; that means, the previous or the previous kind of set up and the kind of future set ups through which actually we can analyze the problem more accurately you know as per the particular requirement, and the kind of best management needs.

(Refer Slide Time: 06:01)



Scope of Data Mining

Some common approaches to data mining

- *Data Exploration and Reduction*
  - identify groups in which elements are similar
- *Classification*
  - analyze data to predict how to classify new elements
- *Association*
  - analyze data to identify natural associations
- *Cause-and-effect Modeling*
  - develop analytic models to describe relationships (e.g.; regression)

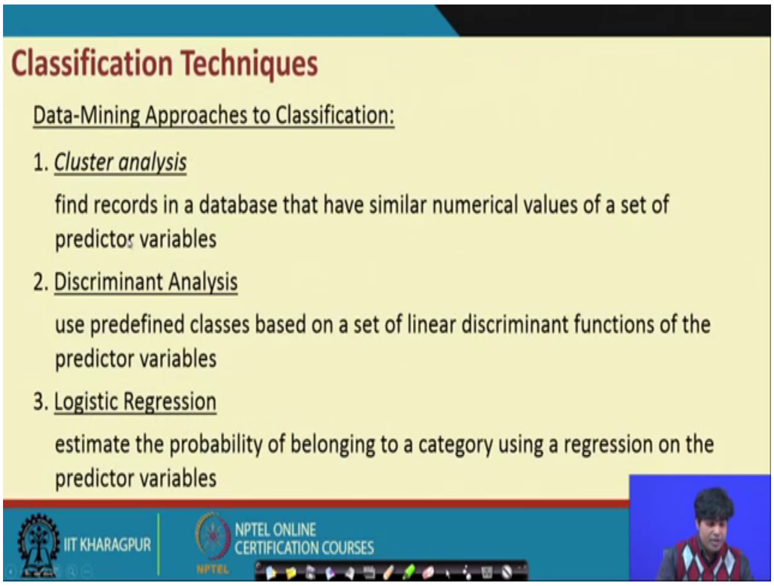
So, likewise so, we have actually kind of system. And so, far as scope is concerned that to in a in the data mining. So, what I have already mentioned. So, we can actually go for data exploration; that means, more visualize and more insights to in the data and sometimes we will go for data reduction process.

In fact, there is a kind of technique called as factor analysis through which the large set of variables can be transferred into a kind of you know, pools through which you can actually have a better clusters then better structures to understand the problems and to go for the kind of predictions. But in this particular lectures we typically go for cluster analysis to with a particular set up we develop a kind of cluster to understand more accurately, the data more accurately the problem, and then the kind of prediction requirement.

So, we have data exploration data reductions data classifications, then association cause and effect modeling. More or less we have already you know we have already covered all these things little bit high and low depending upon a kind of items which we have discussed earlier.

And in this case just we connect since we are discussing actually classification clusters in this particular you know, unit starting with neural network then random forest support vector machines. So, in the similar clusters we are just connecting a technique called as cluster analysis. Then we like to know how the problems can be analyzed more accurately more you know, in a interesting way by developing a system. And developing a kind of classification through which you can understand the issue more accurately and then go for the kind of management requirement.

(Refer Slide Time: 07:54)



Classification Techniques

Data-Mining Approaches to Classification:

1. *Cluster analysis*

   find records in a database that have similar numerical values of a set of predictor variables

2. Discriminant Analysis

   use predefined classes based on a set of linear discriminant functions of the predictor variables

3. Logistic Regression

   estimate the probability of belonging to a category using a regression on the predictor variables

And ha in fact, in the in the kind of classification techniques, particularly the things are very connected with cluster analysis, discriminant analysis and logistic kind of regressions. means the problems which we have already discussed here exploration classification; that means, see here we what we have actually here. So, this is actually one kind of package the kind of you know. So, data exploration and data reductions, then the classification association cause you know. So, these are all actually lots of you know, tool tools are there lots of tools are here also, lots of tools are here also, lots of tools are here also.

So, technically these are all the base through which actually to visualize the data understand the data get insights from the data, then understand the problems. And then develop a system and structures or develop the kind of algorithms through which actually you can predict the environment as per the particular requirement. And then the decision by default will be very effective. So, comparing these kind of structures, if you go little bit more. So, then you will get to know the that in the kind of classification structure, and the counter parts. So what we have actually you know connecting here; the cluster analysis discriminant analysis logistic regressions.

So, that means, the these are the items so, where actually a means these items can be actually connected and with their. In fact, in a kind of similar kind of techniques pools, or techniques basket, where we have a kind of problems, where more number of variables are there having more number of data.

And within the more number of variables, we have some kind of quantitative variables some kind of qualitative variables. And then how we can actually recognize this particular problems depending upon a particular response variable whether it is qualitative in nature, whether it is in a kind of quantitative in nature then, and develop this system through which you will do the predictions as per the particular problem requirement.

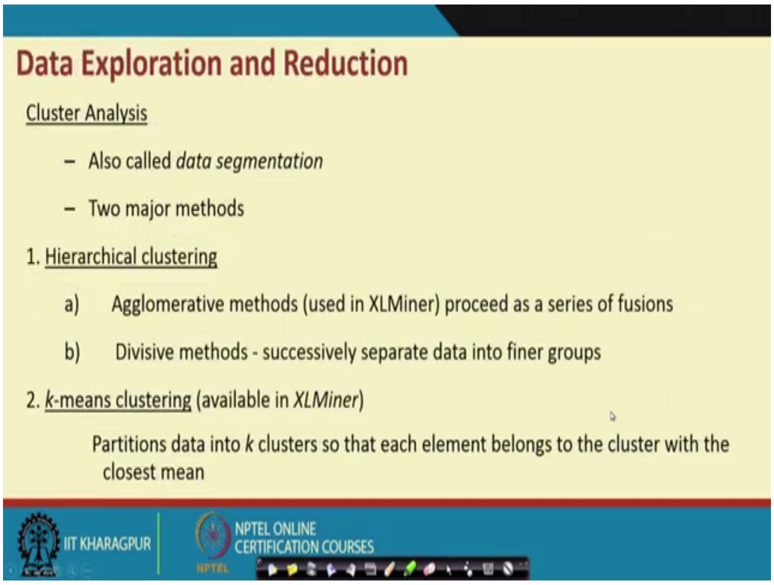
And in this context, cluster analysis discriminant analysis and logistic analysis, logistic regressions are in a kind of more highlighted pictures right. And in fact, in the previous kind of lectures we have already connected or we have already discussed the kind of discriminant analysis, logistic regressions. In fact, in the that is called as a qualitative response regression modeling.

In fact, we have discussed binary stress models that is the linear probability model, which is actually more or less similar to the kind of discriminant analysis, then we have we have discussed logistic regressions followed by logistic functions then similar to we have discussed actually the kind of normal distribution followed by normal distribution that is profit model. And again in the similar scenario we are actually now highlighting the kind of you know structure called as cluster analysis. This is again the kind of classification structure. And through which you will analyze the understand the problem and analyze the problem as per the particular requirement.

Here the means these are the items or these are the techniques predictive analytics tools or techniques, where some of the variables you know, will be kind of qualitative and may be binary in structure or categorical in structure. Then depending upon the kind of in data set and the kind of variables. So, you first check the kind of structure then you can pick up a particular technique through which you can understand the data visualize the data properly.

Then may be go for some kind of reduction mechanism. And then a built a kind of system through which you can do the better kind of prediction better kind of set ups through which you can actually analyze the problem, and come with effective management decision.

(Refer Slide Time: 12:02)



Data Exploration and Reduction

Cluster Analysis

- Also called *data segmentation*
- Two major methods

1. Hierarchical clustering
   a) Agglomerative methods (used in XLMiner) proceed as a series of fusions
   b) Divisive methods - successively separate data into finer groups
2. *k*-means clustering (available in *XLMiner*)
   Partitions data into *k* clusters so that each element belongs to the cluster with the closest mean

And in fact, since our highlights of this particular lectures though we have already discussed discriminant analysis logistic structure that is qualitative response modeling.

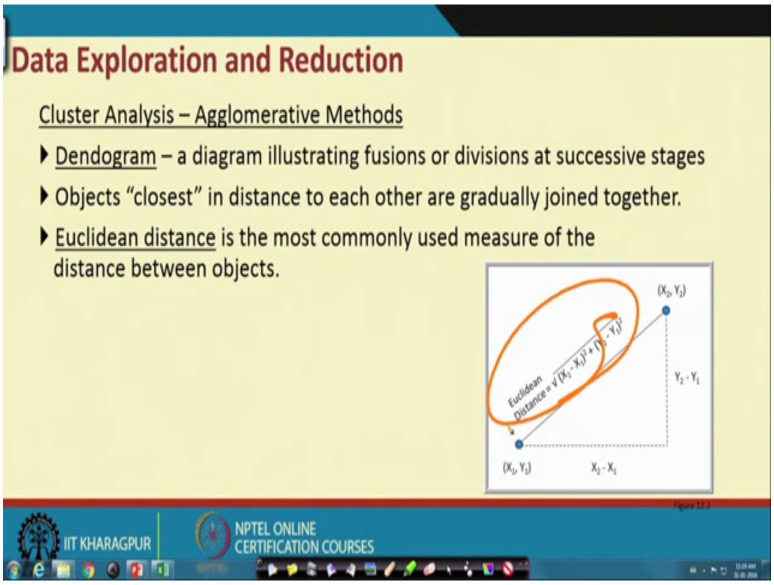
And in this case, we typically highlight or give high stress on cluster analysis. when you talk about the cluster analysis; that means, with a kind of large variables, and large data set we like to prepare a kind of cluster where we can bring the kind of kind of link or structure, where it is a kind of homogenous kind of set ups through which you can do the better prediction, and the kind of better forecasting. And in fact, in the in the kind of you know classification structure that too the particular kind of structure called as cluster analysis.

So, here we have 2 different mechanisms through which you can understand this particular you know predictive analytics tools. And then connect a problems, and check how it is very effective for the management requirement or business requirement. So, there are 2 2 different structures through which we will do the clustering. And one is called as a hierarchical clustering, and then k means clustering. And in the hierarchical clustering's, we have a 2 different methods altogether.

So, agglomerative methods and divisive methods through which you can do the kind of structuring, and then we will understand the system and then find out a kind of set up through which you will do the predictions. So, what I will do? So, I will give you little bit exposure about the hierarchical clustering, and k means clustering.

And then we pick up a kind of problem, then we connect with the both the techniques, and then we see how this this particular technique can be very effective to understand the data understand the kind of problems. And then come with a kind of structures where we can do the effective predictions, and effective kind of forecasting as per the particular management requirement, and the kind of business requirement.

(Refer Slide Time: 14:15)



Data Exploration and Reduction

Cluster Analysis – Agglomerative Methods

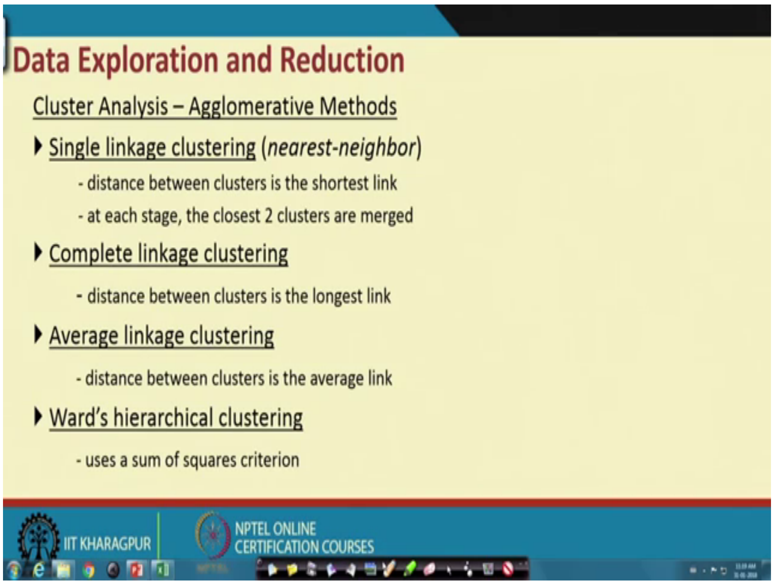
- Dendogram – a diagram illustrating fusions or divisions at successive stages
- Objects "closest" in distance to each other are gradually joined together.
- Euclidean distance is the most commonly used measure of the distance between objects.

(X_2, Y_2)
Euclidean Distance = √(X_2 - X_1)^2 + (Y_2 - Y_1)^2
Y_2 - Y_1
(X_1, Y_1)
X_2 - X_1
Figure 12.2
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, let us come with first the kind of clustering with you know, the kind of structure which we called as hierarchical clustering. And in the hierarchical clustering, particularly what we have mentioned earlier so, we have a 2-different set ups altogether, agglomerative and divisive. And in the case of first methods; so, there are 2 items through which you can analyze and to understand the system that is called as dendogram.

And then Euclidean distance and it is a kind of you know, graphical visualization through which you can understand the particular structure, and then do the set ups for the prediction and forecasting. So, this is what the Euclidean distance and in fact it is all actually what we can say that you know.

(Refer Slide Time: 15:00)



Data Exploration and Reduction

Cluster Analysis – Agglomerative Methods

- Single linkage clustering (*nearest-neighbor*)
  - distance between clusters is the shortest link
  - at each stage, the closest 2 clusters are merged
- Complete linkage clustering
  - distance between clusters is the longest link
- Average linkage clustering
  - distance between clusters is the average link
- Ward's hierarchical clustering
  - uses a sum of squares criterion

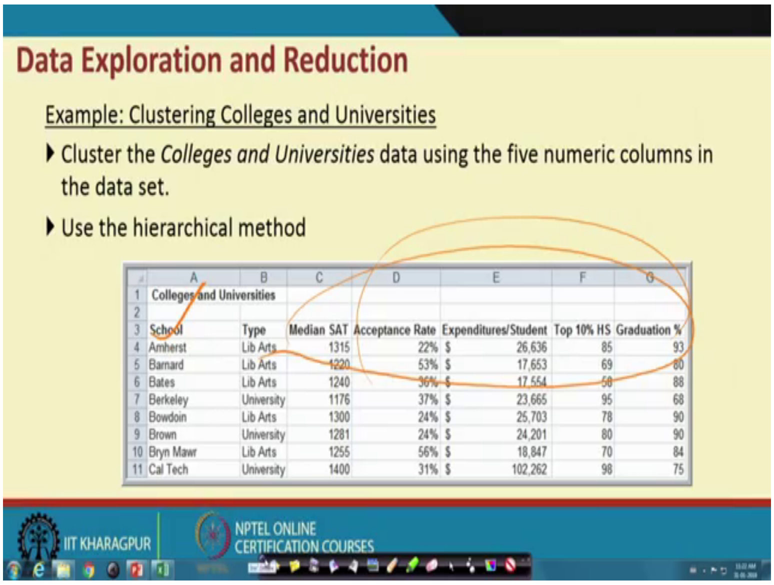
It is a kind of mathematics through which we can understand the particular structure. And this is what the Euclidean distance, and this is what the particular structure. And I will let the dendogram structure through which again you can analyze or you can understand the particular structure. So, this is what simply through mathematics, we can understand the Euclidean distance. And then we use this particular structure through which you can do the better clustering to understand the problems, and the develop the set ups for future prediction and future forecasting's.

So, accordingly so, what I will take you to the particular structure to understand how it is actually effective as per the particular requirement in the clustering. you know, in the cluster analysis particularly this particular method so, we have a single linkage clustering complete linkage clustering average linkage clustering.

So, like we hierarchical clustering. So, we have a different kind of you know models through which you can understand the data, or understand the insights from the data understand the problems then develop the structure through which the manage the problem can be analyzed more accurately as per the management requirement or the kind of business requirement.

So, in any case it is a kind of different means, it is a having a flexibility systems. Within a particular technique, we have a lots of flexibility to analyze the data or get the insights or explore the insights best insights as per the particular, prediction requirement and the kind of management requirement.

(Refer Slide Time: 16:49)



Data Exploration and Reduction

Example: Clustering Colleges and Universities

- Cluster the *Colleges and Universities* data using the five numeric columns in the data set.
- Use the hierarchical method

Colleges and Universities

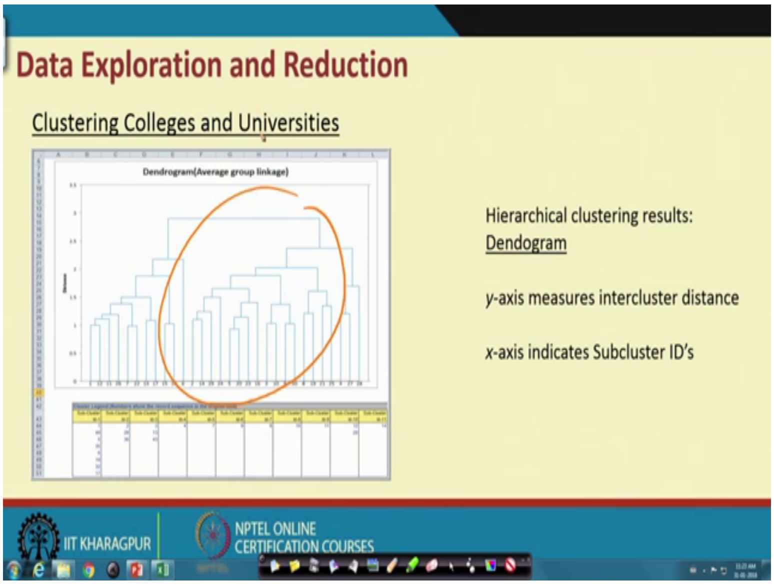
School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
Amherst	Lib Arts	1315	22%	$ 26,636	85	93
Barnard	Lib Arts	1220	53%	$ 17,653	69	80
Bates	Lib Arts	1240	36%	$ 17,554	58	88
Berkeley	University	1176	37%	$ 23,665	95	68
Bowdoin	Lib Arts	1300	24%	$ 25,703	78	90
Brown	University	1281	24%	$ 24,201	80	90
Bryn Mawr	Lib Arts	1255	56%	$ 18,847	70	84
Cal Tech	University	1400	31%	$ 102,262	98	75

So, accordingly s, let us start with a simple examples, and let us say this is a kind of college and university data. And we like to classify the college and university with respect to several attributes in this particular examples we have here’s schools, and then different types schools and colleges. And then we have a different attributes like radiance curve or acceptance rate expenditures students. Or this particular problem we have we have already discussed with a different technique context and problem context. But same problem we are now putting here in the clustering analysis.

Now, here idea is that that with respect to all these attributes, we have to define different clusters, then means what is the management requirement that once you develop different clusters as per the particular requirement, then the policy structure the strategy structure will be accordingly different. And as a result, the or in the entire management system will be very effective and efficient if you develop different clusters. And then plan you know plan accordingly for a particular clusters instead of you know, thinking about to or planning something in a kind of aggregate level.

So, in that context cluster analysis and the data mining package will be very effective for the management requirement, and for the kind of management decisions. So, likewise we have actually. So, let us see here; so, this is what the problem, and then what will you go so, if you analyze.

(Refer Slide Time: 18:28)



Data Exploration and Reduction

Clustering Colleges and Universities

Dendrogram(Average group linkage)

Hierarchical clustering results: Dendogram

y-axis measures intercluster distance

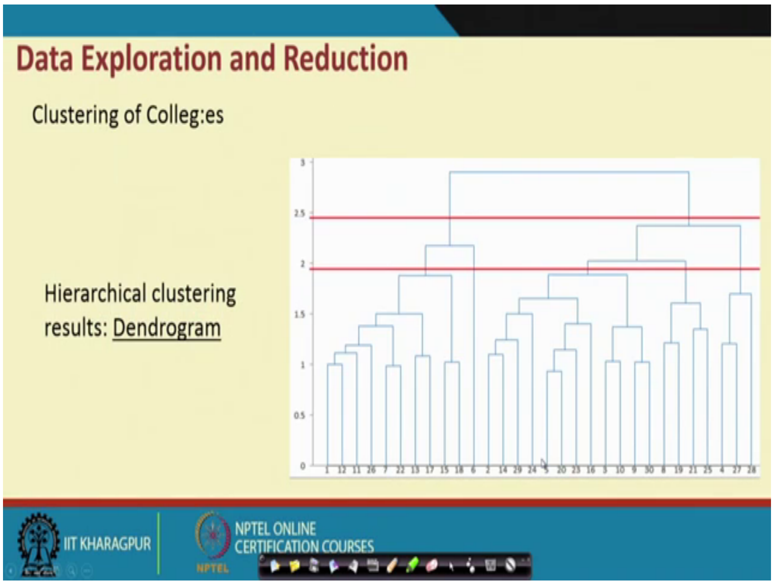
x-axis indicates Subcluster ID's

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, this is what the dendograms, the dendograms structure will be like it is like decision tree structures like what we have already discussed in the case of you know, random forest, and through which how we will connect properly with these attributes, then we will classify the particular structure. And develop the kind of clustering through which you can analyze the problem.

So, this is a simple look about the dendograms through which clustering can be done with respect to college and university, and that too with respect to different attributes. So, I will take you to a kind of software's, then I will highlight with a problems, and then I will show you how the dendogram can be coming and classifying the things or the clustering the things as per the particular business requirement or the problem requirement.

(Refer Slide Time: 19:21)



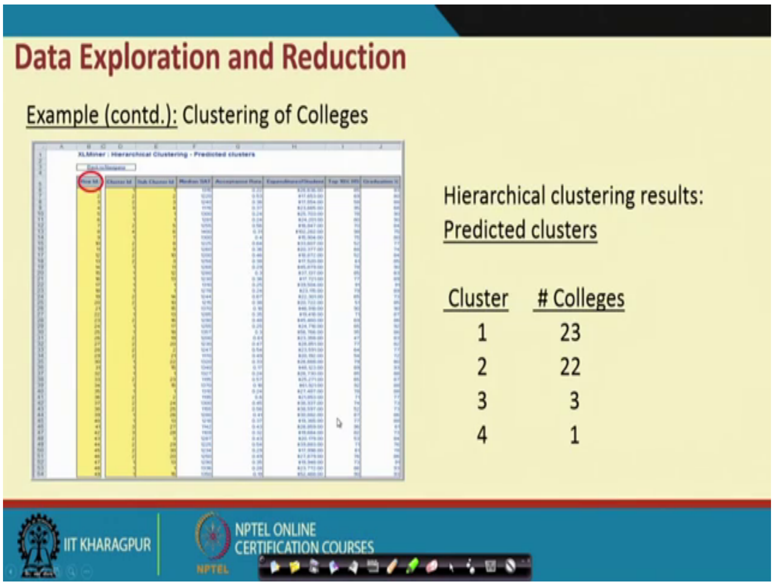
Data Exploration and Reduction

Clustering of Colleg:es

Hierarchical clustering results: Dendrogram

So, then accordingly so, this is another kind of visualization; that means, we have a flexibility system now, if you choose certain dynamics in the kind of structure, then you will get different kind of result then finally, we have a plenty of different structure. Or models or set ups, then we pick up a particular set up which is very effective for the particular problem, and the kind of management requirement.

(Refer Slide Time: 19:45)



Data Exploration and Reduction

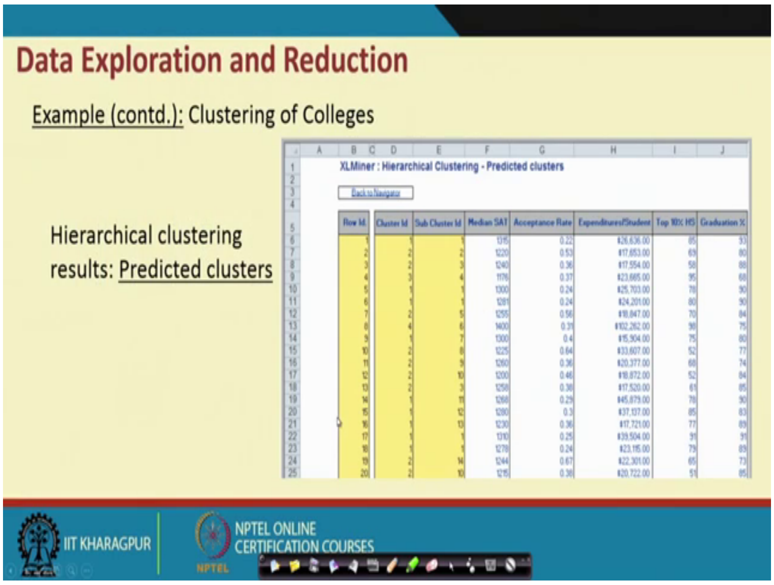
Example (contd.): Clustering of Colleges

Hierarchical clustering results: Predicted clusters

Cluster	# Colleges
1	23
2	22
3	3
4	1

XLMiner : Hierarchical Clustering - Predicted clusters
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

(Refer Slide Time: 19:46)



Data Exploration and Reduction

Example (contd.): Clustering of Colleges

Hierarchical clustering results: Predicted clusters

XLMiner : Hierarchical Clustering - Predicted clusters

Back to Navigator

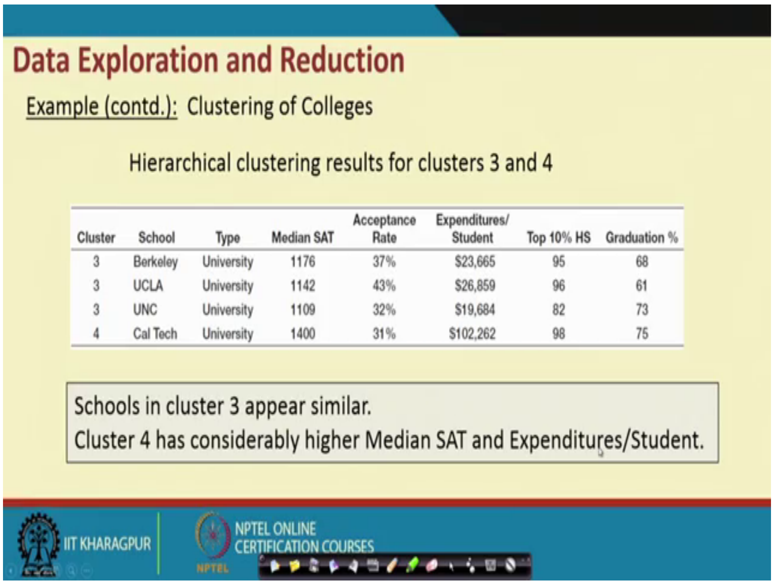
Row Id.	Cluster Id	Sub Cluster Id	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
1	1	1	1315	0.22	$26,636.00	85	93
2	2	2	1220	0.53	$17,653.00	69	80
3	2	3	1240	0.36	$17,554.00	58	88
4	3	4	1176	0.37	$23,665.00	95	68
5	1	1	1300	0.24	$25,703.00	78	90
6	1	1	1281	0.24	$24,201.00	80	90
7	2	5	1255	0.56	$18,847.00	70	84
8	4	6	1400	0.31	$102,262.00	98	75
9	1	7	1300	0.4	$15,904.00	75	80
10	2	8	1225	0.64	$33,607.00	52	77
11	2	9	1260	0.36	$20,377.00	68	74
12	2	10	1200	0.46	$18,872.00	52	84
13	2	3	1258	0.38	$17,520.00	61	85
14	1	11	1268	0.29	$45,879.00	78	90
15	1	12	1280	0.3	$37,137.00	85	83
16	1	13	1230	0.36	$17,721.00	77	89
17	1	1	1310	0.25	$39,504.00	91	91
18	1	1	1278	0.24	$23,115.00	79	89
19	2	14	1244	0.67	$22,301.00	65	73
20	2	10	1215	0.38	$20,722.00	51	85

So, these are all data set, and through which actually we analyze the problem. And then finally, we have in this case in the particular analysis we have 4 different clusters 1, 2, 3, 4. Depending upon the kind of college and the university, and then and the kind of attributes.

So, here in this 4 clusters. So, the first clusters is having 23 colleges, and second cluster is having 22 colleges, third cluster is having 3 colleges, and 4th cluster is having one college one college. So, that means, technically now, so far as management requirement is concerned management decision is concerned. So then effective kind of strategy or effective kind of policy can be implemented differently instead of putting the general kind of strategy or general kind of policy.

So, in that context cluster analysis is always at harsh labels to understand the system. And predict the kind of system as per the requirement.

(Refer Slide Time: 20:45)



Data Exploration and Reduction

Example (contd.): Clustering of Colleges

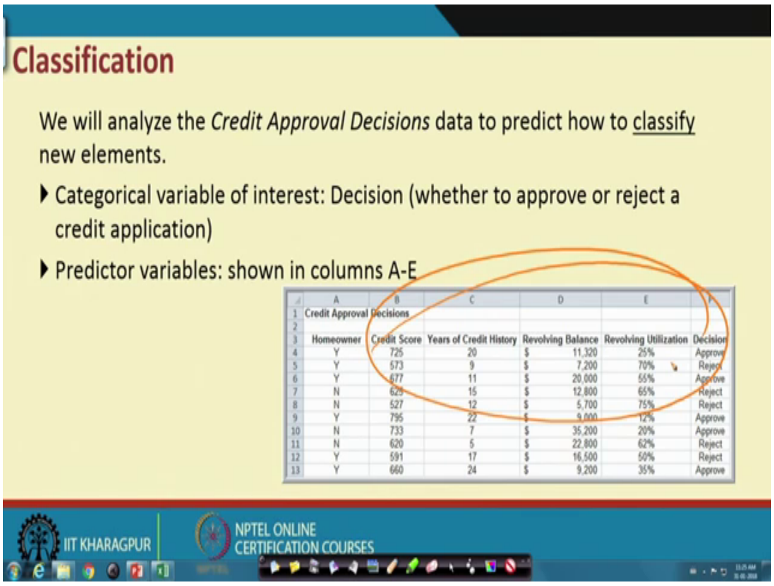
Hierarchical clustering results for clusters 3 and 4

Cluster	School	Type	Median SAT	Acceptance Rate	Expenditures/ Student	Top 10% HS	Graduation %
3	Berkeley	University	1176	37%	$23,665	95	68
3	UCLA	University	1142	43%	$26,859	96	61
3	UNC	University	1109	32%	$19,684	82	73
4	Cal Tech	University	1400	31%	$102,262	98	75

Schools in cluster 3 appear similar.
Cluster 4 has considerably higher Median SAT and Expenditures/Student.

So now with this particular data so, we have actually this classification, these are in the cluster 3. So, these are university is coming. And this is what the attributes through which you do the kind of classifications.

(Refer Slide Time: 20:59)



Classification

We will analyze the *Credit Approval Decisions* data to predict how to classify new elements.

- Categorical variable of interest: Decision (whether to approve or reject a credit application)
- Predictor variables: shown in columns A-E

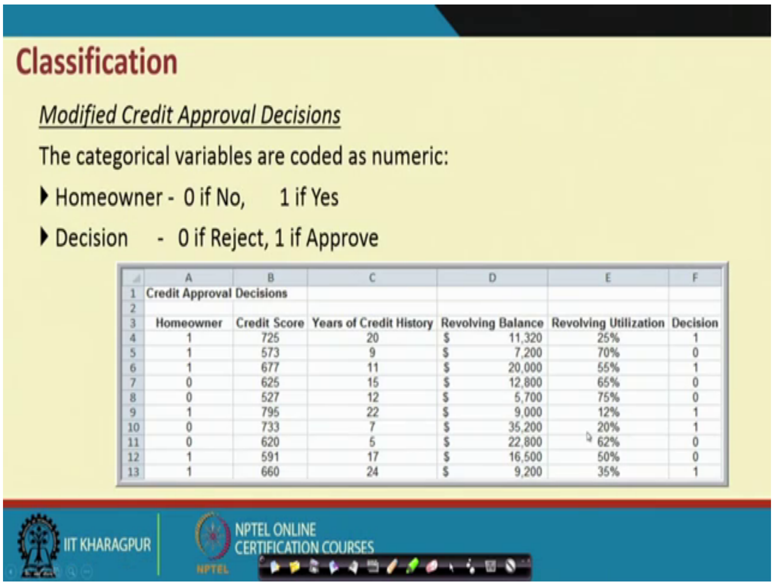
	A	B	C	D	E	F
1	Credit Approval Decisions					
2						
3	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
4	Y	725	20	$ 11,320	25%	Approve
5	Y	573	9	$ 7,200	70%	Reject
6	Y	677	11	$ 20,000	55%	Approve
7	N	625	15	$ 12,800	65%	Reject
8	N	527	12	$ 5,700	75%	Reject
9	Y	795	22	$ 9,000	12%	Approve
10	N	733	7	$ 35,200	20%	Approve
11	N	620	5	$ 22,800	62%	Reject
12	Y	591	17	$ 16,500	50%	Reject
13	Y	660	24	$ 9,200	35%	Approve

So, we will we will we will again analyze with with a kind of software's, and this another kind of examples, this is a credit approval case.

So, credit approved and not approved like in the case of discriminant analysis logistic regression. So, this is also similar kind of structure. So, here so, whether credit approved or not approved, and that too that too exclusively depends upon several factors and several attributes. So, in this kind of example we have several factors here. So, these are all factors credit approved yes or no. Then we have several factors and through which we have take the decision so; that means, we have actually data set.

And then we have to connect a model then after that we can analyze that how a this kind of effective structures, and that too how is the kind of decision about the credit approval with respect to these indicators. So now, now with the help of these particular structures we can actually get better insights into understand the particular technique.

(Refer Slide Time: 22:04)



Classification

Modified Credit Approval Decisions

The categorical variables are coded as numeric:

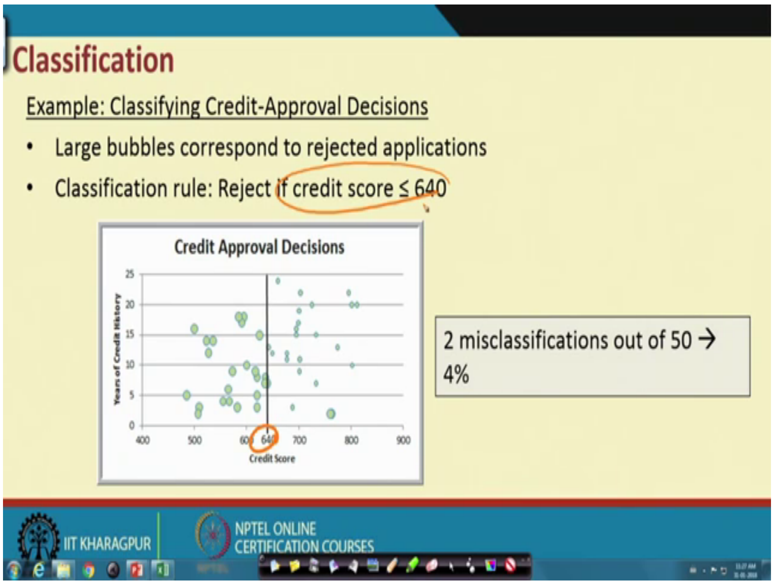
- Homeowner - 0 if No, 1 if Yes
- Decision - 0 if Reject, 1 if Approve

Credit Approval Decisions

Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
1	725	20	$ 11,320	25%	1
1	573	9	$ 7,200	70%	0
1	677	11	$ 20,000	55%	1
0	625	15	$ 12,800	65%	0
0	527	12	$ 5,700	75%	0
1	795	22	$ 9,000	12%	1
0	733	7	$ 35,200	20%	1
0	620	5	$ 22,800	62%	0
1	591	17	$ 16,500	50%	0
1	660	24	$ 9,200	35%	1

So, this is how attributes, that too in the independent clusters.

(Refer Slide Time: 22:08)



Classification

Example: Classifying Credit-Approval Decisions

- Large bubbles correspond to rejected applications
- Classification rule: Reject if credit score ≤ 640

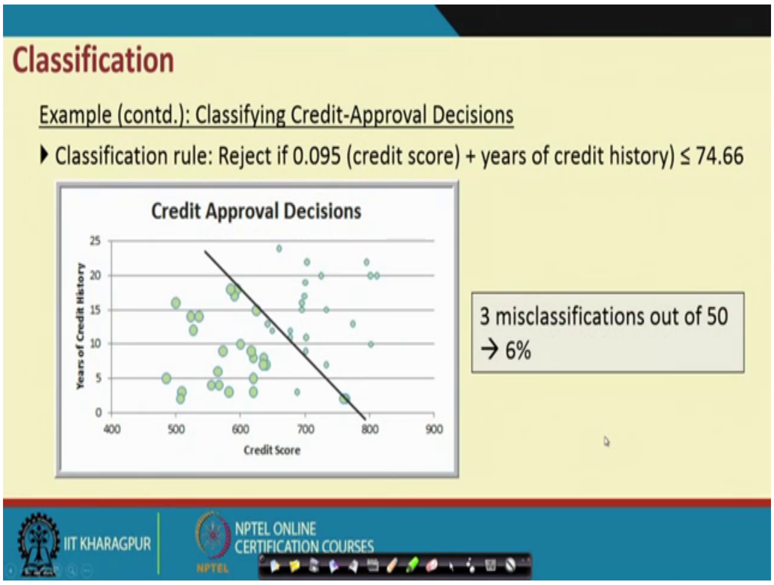
2 misclassifications out of 50 → 4%

Now, having a kind of classification structures, let us say this is bar and through which. So, we have a kind of means if you specify a particular attributes, let us say here in this case it is a 640, then on the basis of that you know, you can give a kind of you know, structure whether to accept and reject. So now, having the kind of bar fixation.

So, these are the kind of means data set which are actually approved and the another group of data vary is not; that means, it is a kind of classifier you find out a kind of kind of structure, through which entire system will be classified into 2 different structure, and then you will analyze as per the particular requirement.

So, that means, it is a kind of interesting structure, through which actually you can develop this, another way of classification.

(Refer Slide Time: 23:00)



Classification

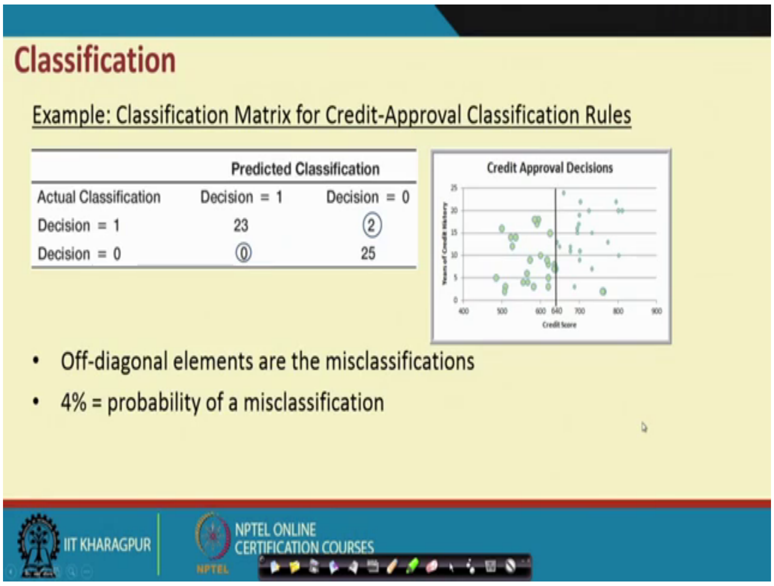
Example (contd.): Classifying Credit-Approval Decisions

- Classification rule: Reject if 0.095 (credit score) + years of credit history) ≤ 74.66

3 misclassifications out of 50
→ 6%

That means, if you change the dynamics at this particular, systems then you will find the particular kind of classifier through which you can understand the data then predict the environment as per the particular business requirement.

(Refer Slide Time: 23:12)



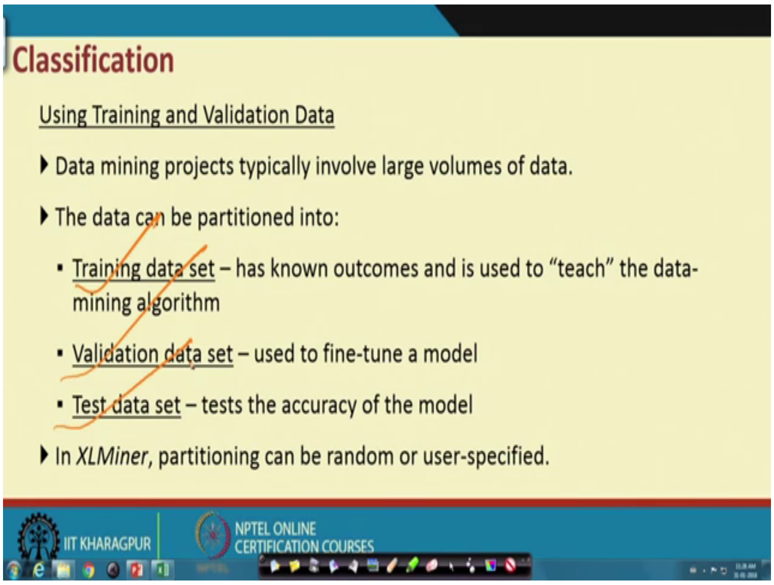
Classification

Example: Classification Matrix for Credit-Approval Classification Rules

	Predicted Classification	
Actual Classification	Decision = 1	Decision = 0
Decision = 1	23	2
Decision = 0	0	25

- Off-diagonal elements are the misclassifications
- 4% = probability of a misclassification

(Refer Slide Time: 23:14)



Classification

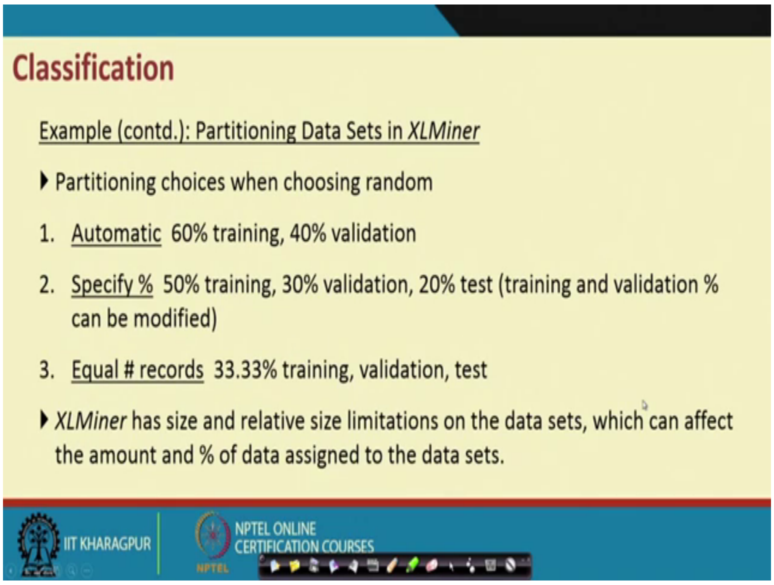
Using Training and Validation Data

- Data mining projects typically involve large volumes of data.
- The data can be partitioned into:
  - Training data set – has known outcomes and is used to "teach" the data-mining algorithm
  - Validation data set – used to fine-tune a model
  - Test data set – tests the accuracy of the model
- In *XLMiner*, partitioning can be random or user-specified.

So, likewise you know, this is what you understand the particular systems and, and then you analyze the problem as per the particular requirement. In the classification structure like we have actually discussed in the case of neural network. In fact, in the random forest and support vector machines; so, every times you can since we have a large data then you know, we first pick up data develop the model, then train the models and then validate and test them finally, fix the kind of structure through which you can go for the effective prediction and effective, forecasting and then effective management decision.

So, it is also same thing here. So, we have here also the training structure validation structure and testing structures. So, this will give you kind of efficient system through which we can actually we will confident that the particular structure which we developed particular everything which we have developed is very effective of the kind of problem, and the kind of management. So, as a result means these are all more or less, same and through which actually you can work and then you analyze.

(Refer Slide Time: 24:25)



Classification

Example (contd.): Partitioning Data Sets in *XLMiner*

- Partitioning choices when choosing random

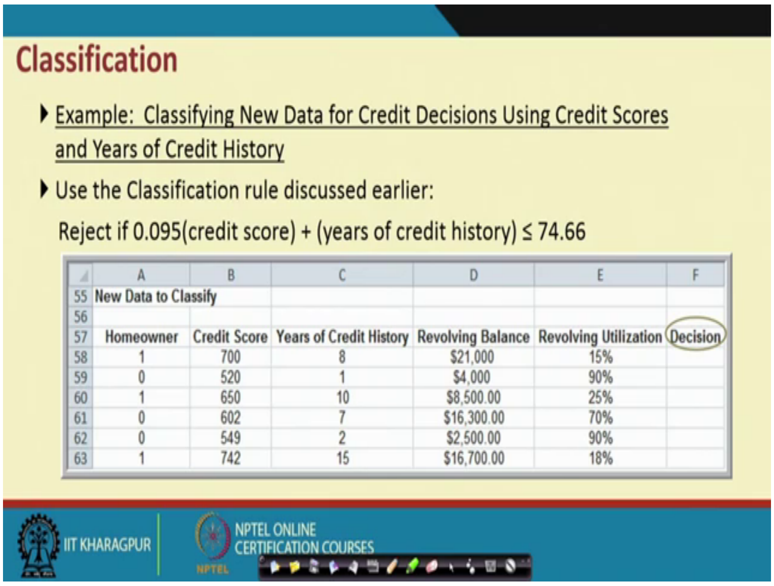
1. Automatic 60% training, 40% validation
2. Specify % 50% training, 30% validation, 20% test (training and validation % can be modified)
3. Equal # records 33.33% training, validation, test

- *XLMiner* has size and relative size limitations on the data sets, which can affect the amount and % of data assigned to the data sets.

So, there is no hard and fast rule what should be the kind of training set, and what should be the validation stress, like we have already highlights this is in the case of neural network random forest and support vector machine.

So, anyway this is what actually; you can actually fix up the particular structure, how many data points you can put for training how much data point we can put for testing. But ultimately, we need a efficient structure through which you can actually look for the efficient prediction efficient forecasting and the kind of management requirement.

(Refer Slide Time: 25:00)



Classification

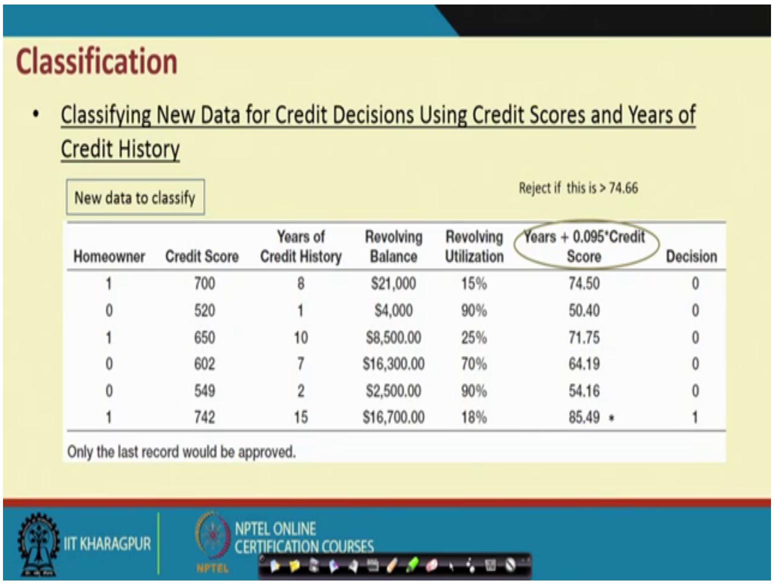
- Example: Classifying New Data for Credit Decisions Using Credit Scores and Years of Credit History
- Use the Classification rule discussed earlier:

Reject if 0.095(credit score) + (years of credit history) $\leq$ 74.66

	A	B	C	D	E	F
55	New Data to Classify					
56						
57	Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Decision
58	1	700	8	$21,000	15%	
59	0	520	1	$4,000	90%	
60	1	650	10	$8,500.00	25%	
61	0	602	7	$16,300.00	70%	
62	0	549	2	$2,500.00	90%	
63	1	742	15	$16,700.00	18%	

And in fact, some classification with the particular data. So now, the decision can means, the kind of model can be different if you again change the particular attribute.

(Refer Slide Time: 25:07)



Classification

- Classifying New Data for Credit Decisions Using Credit Scores and Years of Credit History

New data to classify

Reject if this is > 74.66

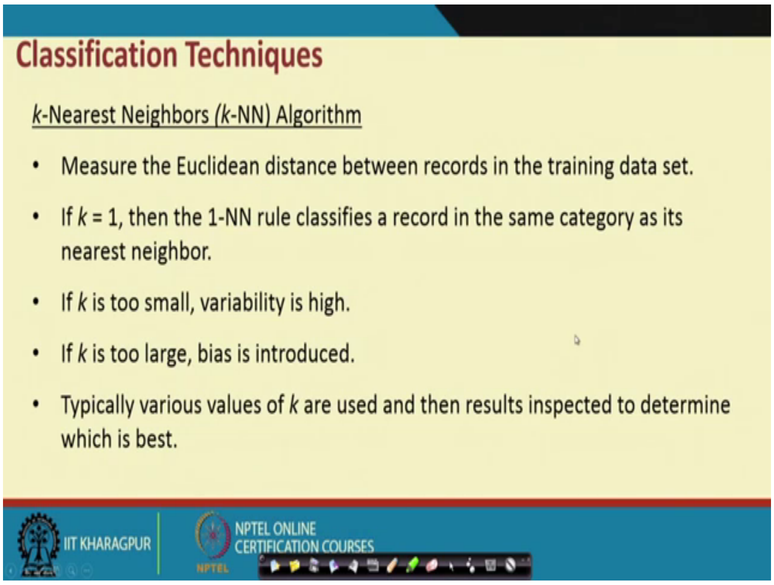
Homeowner	Credit Score	Years of Credit History	Revolving Balance	Revolving Utilization	Years + 0.095*Credit Score	Decision
1	700	8	$21,000	15%	74.50	0
0	520	1	$4,000	90%	50.40	0
1	650	10	$8,500.00	25%	71.75	0
0	602	7	$16,300.00	70%	64.19	0
0	549	2	$2,500.00	90%	54.16	0
1	742	15	$16,700.00	18%	85.49	1

Only the last record would be approved.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Then fix the kind of range, and again you will find 2 different classification group; where it is accepted and the kind of rejected.

(Refer Slide Time: 25:20)



Classification Techniques

k-Nearest Neighbors (*k*-NN) Algorithm

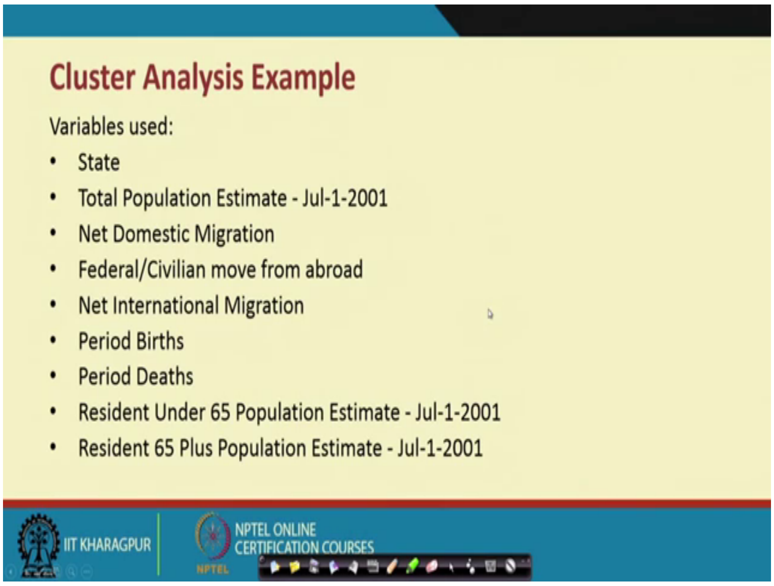
- Measure the Euclidean distance between records in the training data set.
- If *k* = 1, then the 1-NN rule classifies a record in the same category as its nearest neighbor.
- If *k* is too small, variability is high.
- If *k* is too large, bias is introduced.
- Typically various values of *k* are used and then results inspected to determine which is best.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is again similar kind of structuring. And another kind of structure called as k means structuring, and here's it is also more or less same. So, how you have to develop a system, through which you can actually understand the particular insights develop a particular clustering, through which you can analyze the problem more effectively, and develop the kind of structure through which you can actually take a kind of best management decisions. So, in order since we have already discussed 2 kinds of you know, clustering techniques.

So, simple structure then the k mean structures. And what I will do? So, I will take you to a kind of problem, and then I will highlight how clustering can be done, and then what kind of insights or we can observe from the data set and in the kind of problem, the you will find the kind of effective system from the data set, and the kind of variables structure. And then we will think about the kind of management decision, and the kind of management requirement.

(Refer Slide Time: 26:29)



Cluster Analysis Example

Variables used:

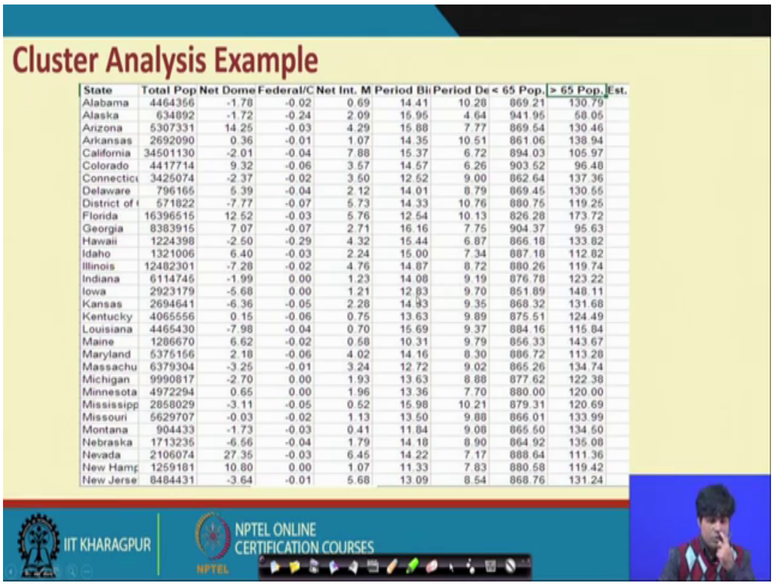
- State
- Total Population Estimate - Jul-1-2001
- Net Domestic Migration
- Federal/Civilian move from abroad
- Net International Migration
- Period Births
- Period Deaths
- Resident Under 65 Population Estimate - Jul-1-2001
- Resident 65 Plus Population Estimate - Jul-1-2001

So, so, what will you do? In the in this kind of to understand the clustering analysis. So, I will take a here the particular problem. So, here variables used in this particular discussion is this state total population, and net domestic migration civilian move, net international migrations periods of birth periods of death then residential populations 60 under 65, and above 65. So, these are the things in the problem basket.

And we have a big data set and then how you have to actually pick up the particular variables, and how you have to clustering. Then that is that needs actually planning and that needs efficient kind of understanding. And of course, it is with respect to management objectives or business objectives. So, that do the kind of classification do the kind of clustering. Without clue or knowing the particular hint you may not go for proper classification or kind of proper clustering.

So, obviously having or large data set and having you know, information about all kind of variables to this particular problem to analyze the problem. So, then you think about to what kind of technique and what kind of structure, or what kind of algorithm you have to develop. So that the particular system can be very effective for the kind of prediction, and the kind of management requirement; so, in order to understand, so what I will do with respect to this particular.

(Refer Slide Time: 28:02)

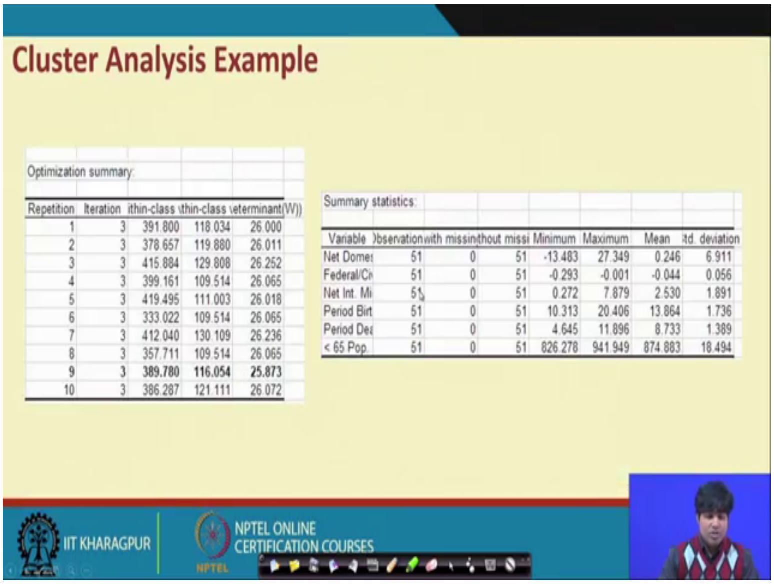


Cluster Analysis Example

State	Total Pop	Net Dome	Federal/C	Net Int. M	Period Bi	Period De	< 65 Pop.	> 65 Pop.	Est.
Alabama	4464356	-1.78	-0.02	0.69	14.41	10.28	869.21	130.79	
Alaska	634892	-1.72	-0.24	2.09	15.95	4.64	941.95	58.05	
Arizona	5307331	14.25	-0.03	4.29	15.88	7.77	869.54	130.46	
Arkansas	2692090	0.36	-0.01	1.07	14.35	10.51	861.06	138.94	
California	34501130	-2.01	-0.04	7.88	15.37	6.72	894.03	105.97	
Colorado	4417714	9.32	-0.06	3.57	14.57	6.26	903.52	96.48	
Connectic	3425074	-2.37	-0.02	3.50	12.52	9.00	862.64	137.36	
Delaware	796165	5.39	-0.04	2.12	14.01	8.79	869.45	130.55	
District of	571822	-7.77	-0.07	5.73	14.33	10.76	880.75	119.25	
Florida	16396515	12.52	-0.03	5.76	12.54	10.13	826.28	173.72	
Georgia	8383915	7.07	-0.07	2.71	16.16	7.75	904.37	95.63	
Hawaii	1224398	-2.50	-0.29	4.32	15.44	6.87	866.18	133.82	
Idaho	1321006	6.40	-0.03	2.24	15.00	7.34	887.18	112.82	
Illinois	12482301	-7.28	-0.02	4.76	14.87	8.72	880.26	119.74	
Indiana	6114745	-1.99	0.00	1.23	14.08	9.19	876.78	123.22	
Iowa	2923179	-5.68	0.00	1.21	12.83	9.70	851.89	148.11	
Kansas	2694641	-6.36	-0.05	2.28	14.93	9.35	868.32	131.68	
Kentucky	4065556	0.15	-0.06	0.75	13.63	9.89	875.51	124.49	
Louisiana	4465430	-7.98	-0.04	0.70	15.69	9.37	884.16	115.84	
Maine	1286670	6.62	-0.02	0.58	10.31	9.79	856.33	143.67	
Maryland	5375156	2.18	-0.06	4.02	14.16	8.30	886.72	113.28	
Massachu	6379304	-3.25	-0.01	3.24	12.72	9.02	865.26	134.74	
Michigan	9990817	-2.70	0.00	1.93	13.63	8.88	877.62	122.38	
Minnesota	4972294	0.65	0.00	1.96	13.36	7.70	880.00	120.00	
Mississipp	2858029	-3.11	-0.05	0.52	15.98	10.21	879.31	120.69	
Missouri	5629707	-0.03	-0.02	1.13	13.50	9.88	866.01	133.99	
Montana	904433	-1.73	-0.03	0.41	11.84	9.08	865.50	134.50	
Nebraska	1713235	-6.56	-0.04	1.79	14.18	8.90	864.92	135.08	
Nevada	2106074	27.35	-0.03	6.45	14.22	7.17	888.64	111.36	
New Hamp	1259181	10.80	0.00	1.07	11.33	7.83	880.58	119.42	
New Jerse	8484431	-3.64	-0.01	5.68	13.09	8.54	868.76	131.24	

So, this is the original data set and so, then we will go for the kind of analysis.

(Refer Slide Time: 28:07)



Cluster Analysis Example

Optimization summary:

Repetition	Iteration	ithin-class	\thin-class	\eterminant(W))
1	3	391.800	118.034	26.000
2	3	378.657	119.880	26.011
3	3	415.884	129.808	26.252
4	3	399.161	109.514	26.065
5	3	419.495	111.003	26.018
6	3	333.022	109.514	26.065
7	3	412.040	130.109	26.236
8	3	357.711	109.514	26.065
9	**3**	**389.780**	**116.054**	**25.873**
10	3	386.287	121.111	26.072

Summary statistics:

Variable	)bservation	with missin	thout missi	Minimum	Maximum	Mean	td. deviation
Net Domes	51	0	51	-13.483	27.349	0.246	6.911
Federal/Ci	51	0	51	-0.293	-0.001	-0.044	0.056
Net Int. Mi	51	0	51	0.272	7.879	2.530	1.891
Period Birt	51	0	51	10.313	20.406	13.864	1.736
Period Dea	51	0	51	4.645	11.896	8.733	1.389
< 65 Pop.	51	0	51	826.278	941.949	874.883	18.494

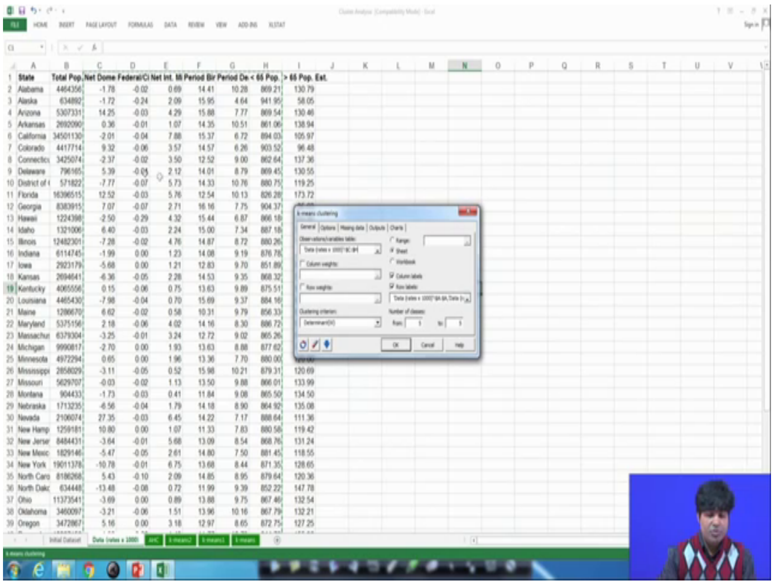
Say, with respect to both the clustering techniques. So, in order to understand. So, I will take you to the kind of excel sheet. So, the same data here.

So, you will find. So, this is a large data set. And having actually you know plenty of information; so whatever I have already highlighted here, this particular structure. So, the same data pool here's, I am actually taking it to the excel sheet. And then, you go to the excel start this same actually software's you can use. In fact, this can be also done through r software. But clustering technique is actually very useful tool, and most of the software's is having this particular module.

And once you understand, then whether you use r software or you know, you use excel start or use SPSS or starta, there is no such kind of you know, problem. Because it is the indication more or less sames, and the outcome through any software. You will derive is also more or less same just you understand the kind of structure, then on understand the problem then the software will help you get the result. So, that we can analyze the problem more effectively, and then you think about the kind of management decision as per the particular business requirement or the kind of problem requirement.

So now understanding the particular problem. So, what will you do here. So, in the kind of clustering analysis, because we have a plenty of techniques same problems can be used by different techniques. And then analyze as per the particular requirement, but since we are in this particular, class we are our discussion is more or less on clustering analysis. So, we like to analyze this problem through clustering analysis, and what we have already discussed that, in the in this particular class we have discussed 2 different clustering technique one is called as agglomerative hierarchical clustering. And then k means clustering. So, let us start with the first k means clustering.

(Refer Slide Time: 30:11)



	A	B	C	D	E	F	G	H	I
1	State	Total Pop.	Net Dome	Federal/Ci	Net Int. Mi	Period Bir	Period De	< 65 Pop.	> 65 Pop. Est.
2	Alabama	4464356	-1.78	-0.02	0.69	14.41	10.28	869.21	130.79
3	Alaska	634892	-1.72	-0.24	2.09	15.95	4.64	941.95	58.05
4	Arizona	5307331	14.25	-0.03	4.29	15.88	7.77	869.54	130.46
5	Arkansas	2692090	0.36	-0.01	1.07	14.35	10.51	861.06	138.94
6	California	34501130	-2.01	-0.04	7.88	15.37	6.72	894.03	105.97
7	Colorado	4417714	9.32	-0.06	3.57	14.57	6.26	903.52	96.48
8	Connecticu	3425074	-2.37	-0.02	3.50	12.52	9.00	862.64	137.36
9	Delaware	796165	5.39	-0.05	2.12	14.01	8.79	869.45	130.55
10	District of (	571822	-7.77	-0.07	5.73	14.33	10.76	880.75	119.25
11	Florida	16396515	12.52	-0.03	5.76	12.54	10.13	826.28	173.72
12	Georgia	8383915	7.07	-0.07	2.71	16.16	7.75	904.37	
13	Hawaii	1224398	-2.50	-0.29	4.32	15.44	6.87	866.18	
14	Idaho	1321006	6.40	-0.03	2.24	15.00	7.34	887.18	
15	Illinois	12482301	-7.28	-0.02	4.76	14.87	8.72	880.26	
16	Indiana	6114745	-1.99	0.00	1.23	14.08	9.19	876.78	
17	Iowa	2923179	-5.68	0.00	1.21	12.83	9.70	851.89	
18	Kansas	2694641	-6.36	-0.05	2.28	14.53	9.35	868.32	
19	Kentucky	4065556	0.15	-0.06	0.75	13.63	9.89	875.51	
20	Louisiana	4465430	-7.98	-0.04	0.70	15.69	9.37	884.16	
21	Maine	1286670	6.62	-0.02	0.58	10.31	9.79	856.33	
22	Maryland	5375156	2.18	-0.06	4.02	14.16	8.30	886.72	
23	Massachu	6379304	-3.25	-0.01	3.24	12.72	9.02	865.26	
24	Michigan	9990817	-2.70	0.00	1.93	13.63	8.88	877.62	
25	Minnesota	4972294	0.65	0.00	1.96	13.36	7.70	880.00	
26	Mississippi	2858029	-3.11	-0.05	0.52	15.98	10.21	879.31	120.69
27	Missouri	5629707	-0.03	-0.02	1.13	13.50	9.88	866.01	133.99
28	Montana	904433	-1.73	-0.03	0.41	11.84	9.08	865.50	134.50
29	Nebraska	1713235	-6.56	-0.04	1.79	14.18	8.90	864.92	135.08
30	Nevada	2106074	27.35	-0.03	6.45	14.22	7.17	888.64	111.36
31	New Hamp	1259181	10.80	0.00	1.07	11.33	7.83	880.58	119.42
32	New Jerse	8484431	-3.64	-0.01	5.68	13.09	8.54	868.76	131.24
33	New Mexic	1829146	-5.47	-0.05	2.61	14.80	7.50	881.45	118.55
34	New York	19011378	-10.78	-0.01	6.75	13.68	8.44	871.35	128.65
35	North Caro	8186268	5.43	-0.10	2.09	14.85	8.95	879.64	120.36
36	North Dakc	634448	-13.48	-0.08	0.72	11.99	9.39	852.22	147.78
37	Ohio	11373541	-3.69	0.00	0.89	13.88	9.75	867.46	132.54
38	Oklahoma	3460097	-3.21	-0.06	1.51	13.96	10.16	867.79	132.21
39	Oregon	3472867	5.16	0.00	3.18	12.97	8.65	872.75	127.25

k-means clustering

General | Options | Missing data | Outputs | Charts

Observations/variables table: 'Data (rates x 1000)'!$C:$I

□ Column weights:

□ Row weights:

Clustering criterion: Determinant(W)

○ Range:
● Sheet
○ Workbook

☑ Column labels
☑ Row labels: 'Data (rates x 1000)'!$A:$A,'Data (r...

Number of classes: From: 1 to: 1

OK | Cancel | Help

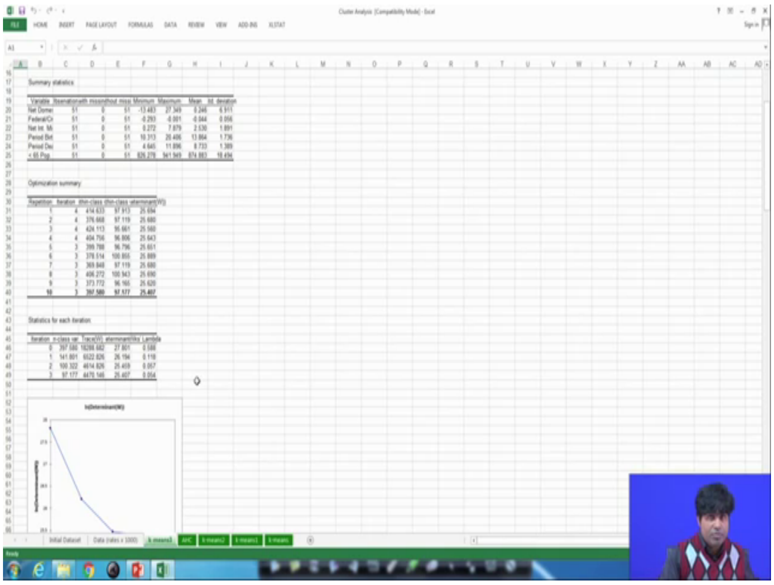
And for that we can click here, and then the input box will be coming like this.

So now it is actually clean. So, what will you do? So, you just indicate how you can do the clustering with respect to all these attributes. So, what will you do? So, you just indicate the kind of variable sheet to these are the variables, which you have given the kind of indications. And then in the row variables also we can clean and give the kind of indications in the row variables. We will give the kind of state indications.

So, that the clustering will be done on the basis of state, and these are the attributes through which you would do the kind of clustering like, which you have already highlighted with you know, with 2 different examples that is with respect to college and university and the kind of credit approval yes and no situation.

Yes, here also same things and after giving all the indications. So, this is how the kind of management objective, once you understand the problem fixing the kind of objective having the data and having the variable information. Just you have to connect, how? as per particular requirement so; that means, you need a complete planning before you come to the software's. And then because software will only be giving you the output through which you can analyze the problem and think about the prediction and think about the kind of insights. But ultimately the planning and the structure which you need to actually operate. So now, after putting all these things we just put and then continue.

(Refer Slide Time: 31:46)



Summary statistics

Variable	bservations	with missin	thout missi	Minimum	Maximum	Mean	td. deviation
Net Domes	51	0	51	-13.483	27.349	0.246	6.911
Federal/Cr	51	0	51	-0.293	-0.001	-0.044	0.056
Net Int. Mi	51	0	51	0.272	7.879	2.530	1.891
Period Bir	51	0	51	10.313	20.406	13.864	1.736
Period Dea	51	0	51	4.645	11.896	8.733	1.389
< 65 Pop.	51	0	51	826.278	941.949	874.883	18.494

Optimization summary

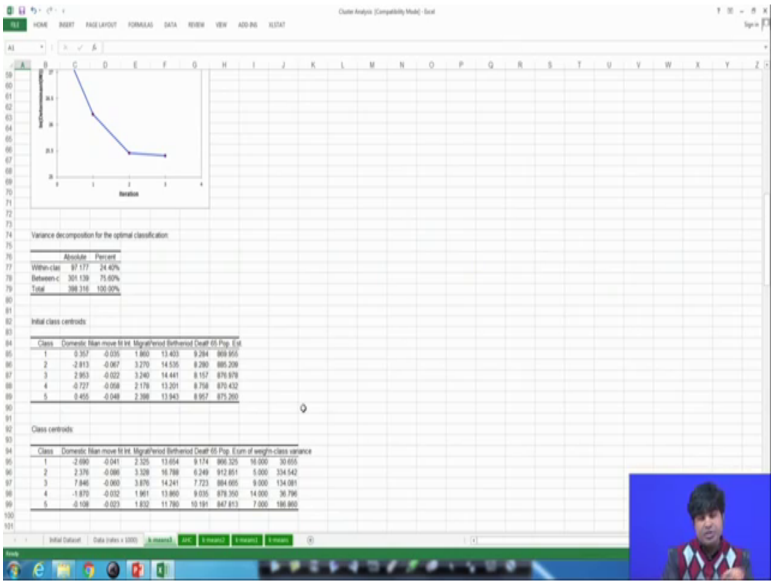
Repetition	Iteration	ithin-class	thin-class	eterminant(W))
1	4	414.633	97.913	25.694
2	4	376.668	97.119	25.680
3	4	424.113	95.661	25.560
4	4	404.756	96.806	25.643
5	3	399.788	96.796	25.651
6	3	378.514	100.855	25.889
7	3	369.848	97.119	25.680
8	3	406.272	100.943	25.690
9	3	373.772	96.165	25.620
10	**3**	**397.580**	**97.177**	**25.407**

Statistics for each iteration

Iteration	n-class var	Trace(W)	eterminant(W)	Wilks' Lambda
0	397.580	10288.682	27.801	0.588
1	141.801	6522.826	26.194	0.118
2	100.322	4614.826	25.459	0.057
3	97.177	4470.146	25.407	0.054

So, this is how the k means clustering's result is coming. So, as usual. So, every techniques you can get a summary statistics, but as descriptive statistic. That is how descriptive analytics is descriptive analytics is the first priority to understand before you go to the predictive analytics and prescriptive analytics. And in fact, in this case we have lots of these are all basic statistics, and this is how the kind of optimization summary and in the statistic of each iteration. These are all k means clustering different class and different iterations, and then again, we will have actually some kind of statistics about the a clustering's.

(Refer Slide Time: 32:27)



Variance decomposition for the optimal classification:

	Absolute	Percent
Within-clas	97.177	24.40%
Between-c	301.139	75.60%
Total	398.316	100.00%

Initial class centroids:

Class	Domestic	Man move fit	Int. Migra	Period Birth	Period Deat	65 Pop. Est
1	0.357	-0.035	1.860	13.403	9.284	869.955
2	-2.813	-0.067	3.270	14.535	8.280	885.209
3	2.953	-0.022	3.240	14.441	8.157	876.978
4	-0.727	-0.058	2.178	13.201	8.758	870.432
5	0.455	-0.048	2.398	13.943	8.957	875.260

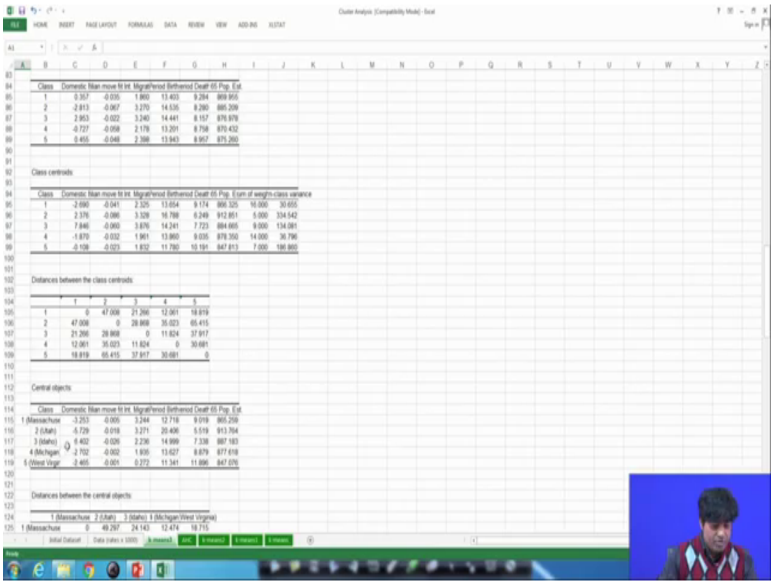
Class centroids:

Class	Domestic	Man move fit	Int. Migra	Period Birth	Period Deat	65 Pop.	Sum of weigh	Within-class variance
1	-2.690	-0.041	2.325	13.654	9.174	866.325	16.000	30.655
2	2.376	-0.086	3.328	16.788	6.249	912.851	5.000	334.542
3	7.846	-0.060	3.876	14.241	7.723	884.665	9.000	134.081
4	-1.870	-0.032	1.961	13.860	9.035	878.350	14.000	36.796
5	-0.108	-0.023	1.832	11.780	10.191	847.813	7.000	186.860

So, this is how the indication through which the iteration can be stopped, because once you start like that. You know, clustering can be done. So now on the graphically; if you and a particular structure is coming like this, so then that will give you the signal that your clustering is done, and that too that is actually effective kind of structures. Once you find all these things, then this is how the kind of summary sheet about the validations, and within the class between the class and between the class 75 percent 76 percent, then this is 24 percent and again. So, we have a lots of classifications with respect to different attributes. So, these are all various classification.

In fact, from this result, you can just guess that manually it is very difficult to done. This kind of you know, clustering and this kind of classification until and unless you connect with a particular software, and understand the particular technique.

(Refer Slide Time: 33:30)



Class	Domestic	Nan move	ft Int. Migra	Period Birth	eriod Deat	65 Pop. Est.
1	0.357	-0.035	1.860	13.403	9.284	869.955
2	-2.813	-0.067	3.270	14.535	8.280	885.209
3	2.953	-0.022	3.240	14.441	8.157	876.978
4	-0.727	-0.058	2.178	13.201	8.758	870.432
5	0.455	-0.048	2.398	13.943	8.957	875.260

Class centroids

Class	Domestic	Nan move	ft Int. Migra	Period Birth	eriod Deat	65 Pop. E	Sum of weigh	n-class variance
1	-2.690	-0.041	2.325	13.654	9.174	866.325	16.000	30.655
2	2.376	-0.086	3.328	16.788	6.249	912.851	5.000	334.542
3	7.846	-0.060	3.876	14.241	7.723	884.665	9.000	134.081
4	-1.870	-0.032	1.961	13.860	9.035	878.350	14.000	36.796
5	-0.108	-0.023	1.832	11.780	10.191	847.813	7.000	186.860

Distances between the class centroids

	1	2	3	4	5
1	0	47.008	21.266	12.061	18.819
2	47.008	0	28.868	35.023	65.415
3	21.266	28.868	0	11.824	37.917
4	12.061	35.023	11.824	0	30.681
5	18.819	65.415	37.917	30.681	0

Central objects

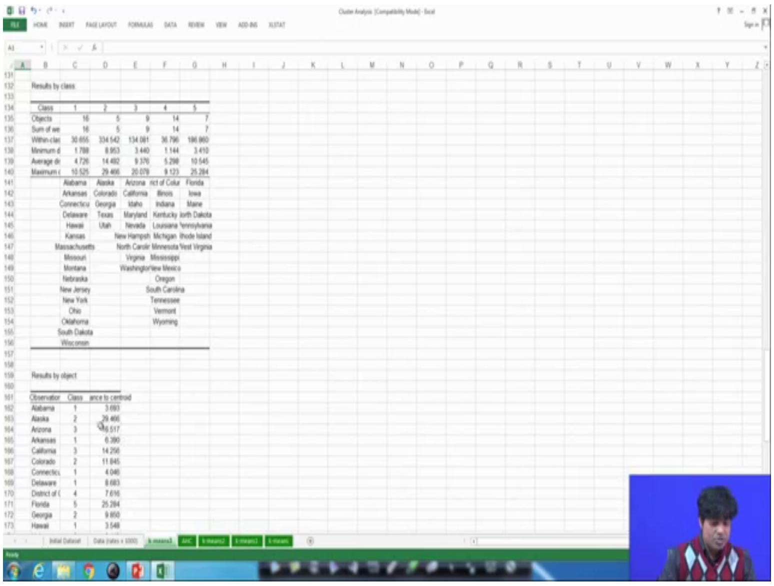
Class	Domestic	Nan move	ft Int. Migra	Period Birth	eriod Deat	65 Pop. Est.
1 (Massachuse	-3.253	-0.005	3.244	12.718	9.019	865.259
2 (Utah)	-5.729	0.018	3.271	20.406	5.519	913.764
3 (Idaho)	6.402	-0.026	2.236	14.999	7.338	887.183
4 (Michigan	-2.702	-0.002	1.935	13.627	8.879	877.618
5 (West Virgi	-2.465	-0.001	0.272	11.341	11.896	847.076

Distances between the central objects

	1 (Massachuse	2 (Utah)	3 (Idaho)	4 (Michigan	West Virginia)
1 (Massachuse	0	49.297	24.143	12.474	18.715

And this is how distance between the class enters. And finally, and these are all different clustering with respect to different attributes.

(Refer Slide Time: 33:40)



Results by class

Class	1	2	3	4	5
Objects	16	5	9	14	7
Sum of we	16	5	9	14	7
Within-clas	30.655	334.542	134.081	36.796	186.860
Minimum d	1.788	8.953	3.440	1.144	3.410
Average di	4.726	14.492	9.376	5.298	10.545
Maximum c	10.525	29.466	20.078	9.123	25.284
	Alabama	Alaska	Arizona	rict of Colur	Florida
	Arkansas	Colorado	California	Illinois	Iowa
	Connecticu	Georgia	Idaho	Indiana	Maine
	Delaware	Texas	Maryland	Kentucky	lorth Dakota
	Hawaii	Utah	Nevada	Louisiana	Pennsylvania
	Kansas		New Hampsh	Michigan	Rhode Island
	Massachusetts		North Carolir	Minnesota	West Virginia
	Missouri		Virginia	Mississippi	
	Montana		Washington	New Mexico	
	Nebraska			Oregon	
	New Jersey			South Carolina	
	New York			Tennessee	
	Ohio			Vermont	
	Oklahoma			Wyoming	
	South Dakota				
	Wisconsin				

Results by object

Observation	Class	ance to centroid
Alabama	1	3.693
Alaska	2	29.466
Arizona	3	16.517
Arkansas	1	6.390
California	3	14.256
Colorado	2	11.845
Connecticu	1	4.046
Delaware	1	8.683
District of C	4	7.616
Florida	5	25.284
Georgia	2	9.850
Hawaii	1	3.548

And by the way you will find plenty of clustering like this, with respect o particular you know, problem and the kind of the kind of cities. And with different attributes like this, this is the class, and then this how the centroid. And so, of course, you know; that means, technically so, this is how the effective clustering can be you know, having to understand the problem, and to understand the structure through which analyze the problems right ok.

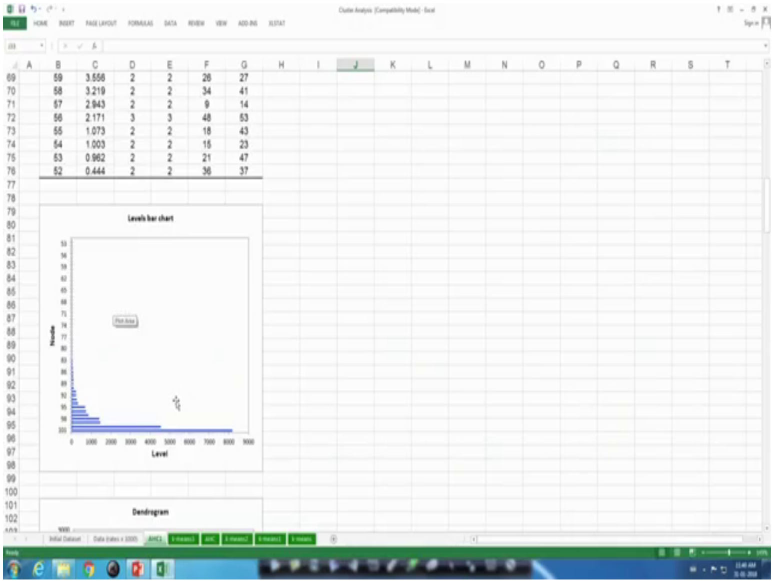
So, what will you do is so, these are the results through which actually you are getting through k means, clustering so; that means, different iterations and different training structure is developed. Then finally, you pick up a particular structure through which you can have actually have a better clustering as per the particular requirement, and then go for the kind of prediction. And the kind of management decision again the same problems can be also analyzed through the other you know, different techniques.

And that too the kind of structure which we have discussed called as dendogram. And the kind of Euclidean distance let us say, go for the particular data set again. So, this is the original data set against you can go to the excel start data analysis, and then this will be under analysing data, and we have already solved this problem through k means clustering.

Now, we can analyze this problem, other mechanism that is agglomerative hierarchical clustering's so; that means, we have discussed k means clustering. And hierarchical clustering k means, clustering actually it is a very different kind of clustering through different iterations. And then final kind of structure through which you can get the better structure or better prediction and against the same structure here. So, we are giving the kind of you know, indication about the variable same the kind of you know, attributes which we have already highlighted here. And then you just put and against continue, you will get similar kind of results like previous case, we have here also some descriptive statistics.

So, in the descriptive statistics. So, these are all various kind of structure, summary sheet, and then the more interesting part of this particular you know, hierarchical clustering is the kind of label bar chart.

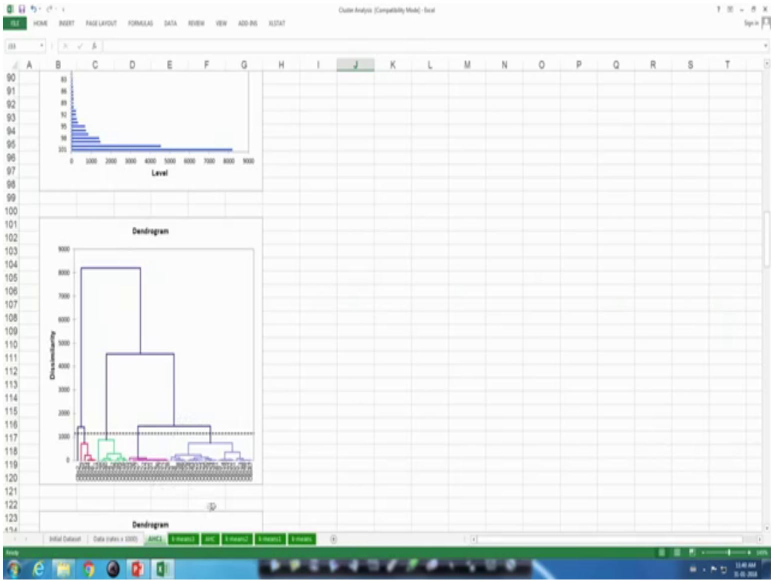
(Refer Slide Time: 35:58)



	A	B	C	D	E	F	G
69		59	3.556	2	2	26	27
70		58	3.219	2	2	34	41
71		57	2.943	2	2	9	14
72		56	2.171	3	3	48	53
73		55	1.073	2	2	18	43
74		54	1.003	2	2	15	23
75		53	0.962	2	2	21	47
76		52	0.444	2	2	36	37

And then the kind of dendogram.

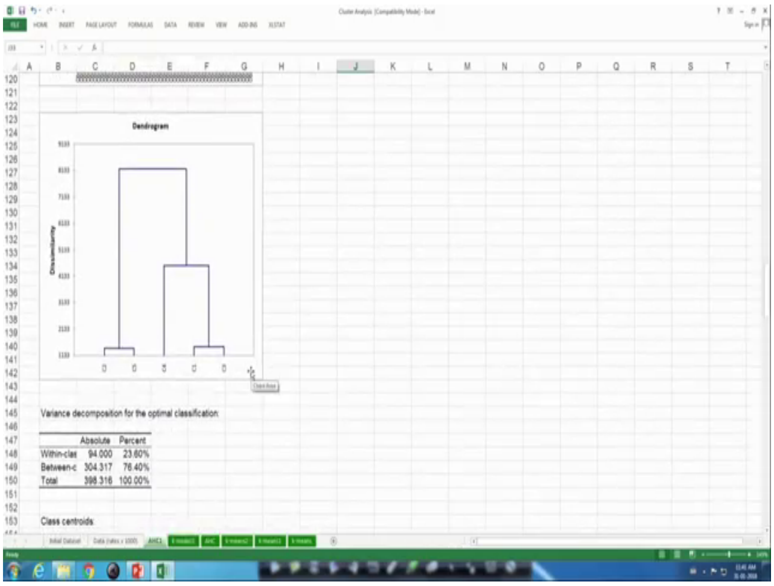
(Refer Slide Time: 36:01)



See here this is a actually the kind of you know, structure through which you can actually clustering the kind of you know, data get understand the insights. And build a kind of classification through which you can do the prediction so; that means, it is a kind of the particular technique is kind of effective tools through which you know, you can better understand the data; visualize the data, and then have a kind of cluster. So, that the management problem can be analyzed more effectively as per the business requirement, and the kind of problem requirement, and do the predictions and do the kind of effective management decision.

So, that is why means it is a very easy to get ok.

(Refer Slide Time: 36:43)



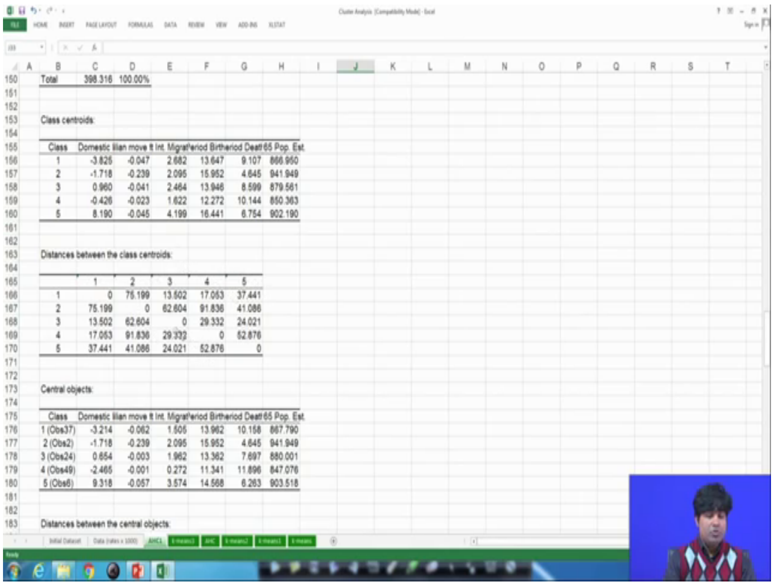
Variance decomposition for the optimal classification:

	Absolute	Percent
Within-clas	94.000	23.60%
Between-c	304.317	76.40%
Total	398.316	100.00%

Class centroids:

So, this is a dendogram another look and these are all like, this is also coming under k means clustering's.

(Refer Slide Time: 36:46)



Total	398.316	100.00%

Class centroids:

Class	Domestic	llian move	ft Int. Migra	Period Birth	Period Deat	65 Pop. Est.
1	-3.825	-0.047	2.682	13.647	9.107	866.950
2	-1.718	-0.239	2.095	15.952	4.645	941.949
3	0.960	-0.041	2.464	13.946	8.599	879.561
4	-0.426	-0.023	1.622	12.272	10.144	850.363
5	8.190	-0.045	4.199	16.441	6.754	902.190

Distances between the class centroids:

	1	2	3	4	5
1	0	75.199	13.502	17.053	37.441
2	75.199	0	62.604	91.836	41.086
3	13.502	62.604	0	29.332	24.021
4	17.053	91.836	29.332	0	52.876
5	37.441	41.086	24.021	52.876	0

Central objects:

Class	Domestic	llian move	ft Int. Migra	Period Birth	Period Deat	65 Pop. Est.
1 (Obs37)	-3.214	-0.062	1.505	13.962	10.158	867.790
2 (Obs2)	-1.718	-0.239	2.095	15.952	4.645	941.949
3 (Obs24)	0.654	-0.003	1.962	13.362	7.697	880.001
4 (Obs49)	-2.465	-0.001	0.272	11.341	11.896	847.076
5 (Obs6)	9.318	-0.057	3.574	14.568	6.263	903.518

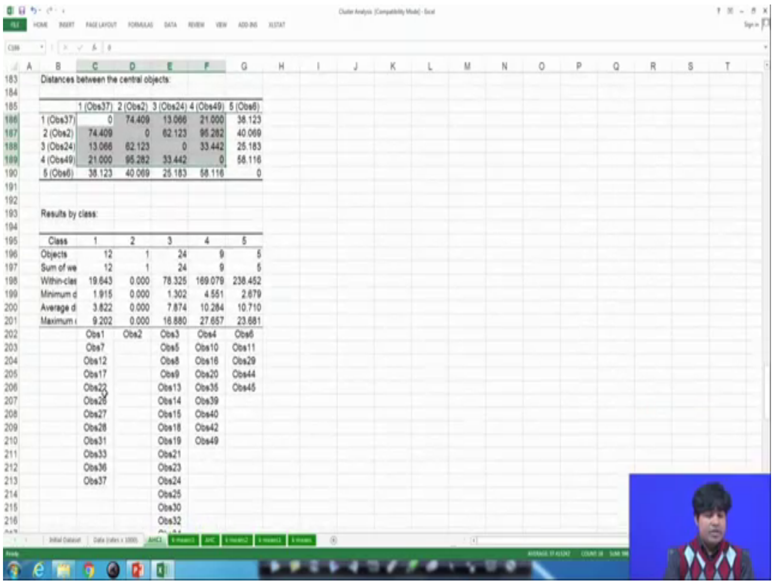
Distances between the central objects:

So, this is how the validation about the means, various variance de composition; that is called as robustness of you know, this particular results. And then finally, there are different you know, class and then these are all with respect to different attributes. And it is more or less same like as the k means clustering. But in the hierarchical clustering, one of the most important thing is the kind of the graphical structure to understand the particular, link and the kind of clustering.

That is the dendogram and then accordingly we can understand, this is what the distance between that is the Euclidean distance, which is actually since we have actually 5 different class. So, we have a actually the kind of matrix here 5 into 5 and; obviously, 1 to 1. So, the distance will be 0 then 2 to 2; obviously, the distance will be 0 again 3 to 3 distance.

So, as a result all the diagonal elements will be having actually 0, figures and others depends upon how the distance is coming actually to a particular, point. So, these are all central objects and that too with different attributes again.

(Refer Slide Time: 37:59)



Distances between the central objects:

	1 (Obs37)	2 (Obs2)	3 (Obs24)	4 (Obs49)	5 (Obs6)
1 (Obs37)	0	74.409	13.066	21.000	38.123
2 (Obs2)	74.409	0	62.123	95.282	40.069
3 (Obs24)	13.066	62.123	0	33.442	25.183
4 (Obs49)	21.000	95.282	33.442	0	58.116
5 (Obs6)	38.123	40.069	25.183	58.116	0

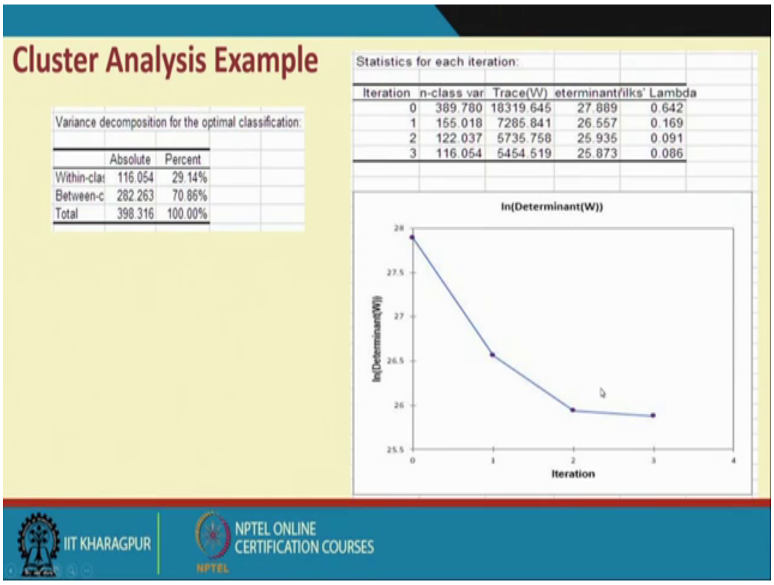
Results by class:

Class	1	2	3	4	5
Objects	12	1	24	9	5
Sum of we	12	1	24	9	5
Within-clas	19.643	0.000	78.325	169.079	238.452
Minimum d	1.915	0.000	1.302	4.551	2.679
Average d	3.822	0.000	7.874	10.284	10.710
Maximum (	9.202	0.000	16.880	27.657	23.681
	Obs1	Obs2	Obs3	Obs4	Obs6
	Obs7		Obs5	Obs10	Obs11
	Obs12		Obs8	Obs16	Obs29
	Obs17		Obs9	Obs20	Obs44
	Obs22		Obs13	Obs35	Obs45
	Obs26		Obs14	Obs39	
	Obs27		Obs15	Obs40	
	Obs28		Obs18	Obs42	
	Obs31		Obs19	Obs49	
	Obs33		Obs21		
	Obs36		Obs23		
	Obs37		Obs24		
			Obs25		
			Obs30		
			Obs32		

So, distance between the central objects again. So, the particular diagonal elements will be 0 as per the Euclidean distance measures, and these are all various clusters class wise. And that too with descriptive statistics, these are all class wise descriptive statistic. And then we can get better to understand now you see here in the third class, we have actually sufficient number of kind of data points. And then the observations, and the as a results you can actually this is more, summary statistic so; that means, it is a more interesting is the kind of this is the kind of classification through which actually, clustering is done and the kind of analysis is actually done and accordingly.

So, this is what the kind of you know, structure.

(Refer Slide Time: 38:49)



Cluster Analysis Example

Variance decomposition for the optimal classification:

	Absolute	Percent
Within-clas	116.054	29.14%
Between-c	282.263	70.86%
Total	398.316	100.00%

Statistics for each iteration:

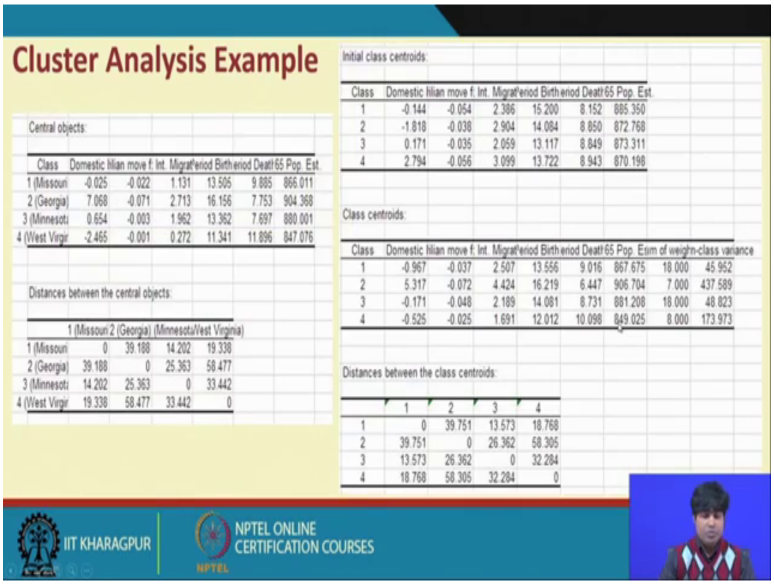
Iteration	n-class var	Trace(W)	eterminant(	ilks' Lambda
0	389.780	18319.645	27.889	0.642
1	155.018	7285.841	26.557	0.169
2	122.037	5735.758	25.935	0.091
3	116.054	5454.519	25.873	0.086

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

And what we have already discussed this is how the graphical understanding about the you know, validations to justify. That you know, this is what the effective clustering through which you can do the kind of you know, understand the particular you know, problem and do the kind of prediction as per the particular you know, business requirement and problem.

(Refer Slide Time: 39:12)



Cluster Analysis Example

Central objects:

Class	Domestic	hilian move f:	Int. Migrat	eriod Birth	eriod Deatl	65 Pop. Est.
1 (Missouri	-0.025	-0.022	1.131	13.506	9.885	866.011
2 (Georgia)	7.068	-0.071	2.713	16.156	7.753	904.368
3 (Minnesota	0.654	-0.003	1.962	13.362	7.697	880.001
4 (West Virgir	-2.465	-0.001	0.272	11.341	11.896	847.076

Distances between the central objects:

	1 (Missouri	2 (Georgia)	(Minnesota	Vest Virginia)
1 (Missouri	0	39.188	14.202	19.338
2 (Georgia)	39.188	0	25.363	58.477
3 (Minnesota	14.202	25.363	0	33.442
4 (West Virgir	19.338	58.477	33.442	0

Initial class centroids:

Class	Domestic	hilian move f:	Int. Migrat	eriod Birth	eriod Deatl	65 Pop. Est.
1	-0.144	-0.054	2.386	15.200	8.152	885.350
2	-1.818	-0.038	2.904	14.084	8.850	872.768
3	0.171	-0.035	2.059	13.117	8.849	873.311
4	2.794	-0.056	3.099	13.722	8.943	870.198

Class centroids:

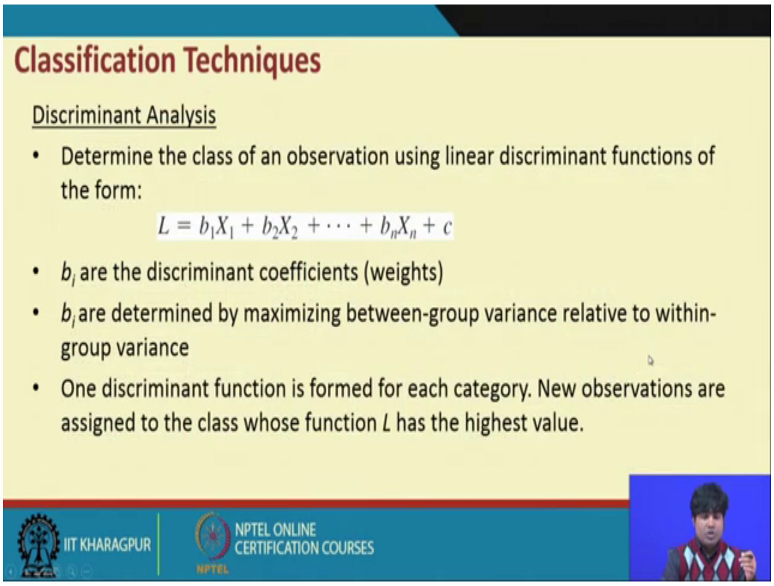
Class	Domestic	hilian move f:	Int. Migrat	eriod Birth	eriod Deatl	65 Pop. E	sim of weigh	n-class variance
1	-0.967	-0.037	2.507	13.556	9.016	867.675	18.000	45.952
2	5.317	-0.072	4.424	16.219	6.447	906.704	7.000	437.589
3	-0.171	-0.048	2.189	14.081	8.731	881.208	18.000	48.823
4	-0.525	-0.025	1.691	12.012	10.098	849.025	8.000	173.973

Distances between the class centroids:

	1	2	3	4
1	0	39.751	13.573	18.768
2	39.751	0	26.362	58.305
3	13.573	26.362	0	32.284
4	18.768	58.305	32.284	0

So, these are the results, which we have already seen from the you know, spread sheet and. And I am just bringing to highlight it this these are the results through which you know, the problem can be analyzed on or can be understand, you can understand better. And through this particular clustering, and that too through k means clustering and through the kind of the kind of hierarchical clustering.

(Refer Slide Time: 39:34)



Classification Techniques

Discriminant Analysis

- Determine the class of an observation using linear discriminant functions of the form:

$$L = b_1X_1 + b_2X_2 + \cdots + b_nX_n + c$$

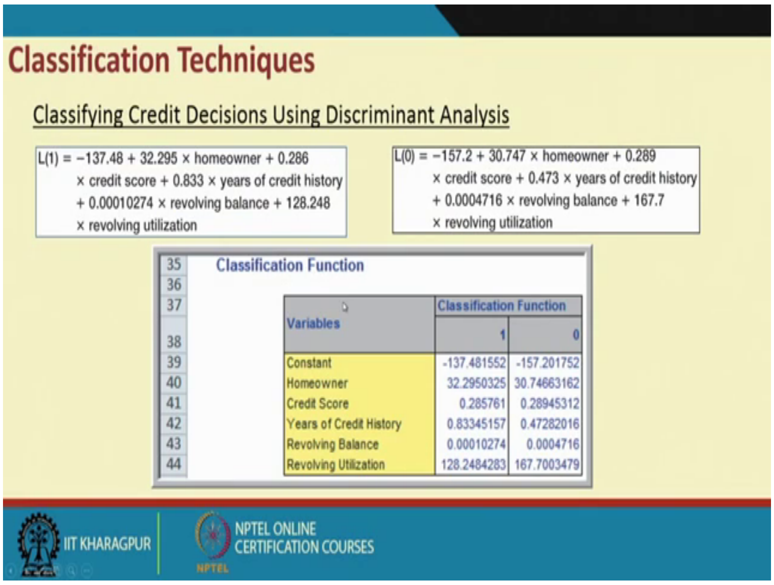
- b_i are the discriminant coefficients (weights)
- b_i are determined by maximizing between-group variance relative to within-group variance
- One discriminant function is formed for each category. New observations are assigned to the class whose function L has the highest value.

And the corresponding to the kind of class classification, means clustering techniques like k means, clustering and hierarchical clustering.

So, there are also some of the classification techniques under the data mining; which can also be very effective, through which you can understand the problems get the insights. And do the kind of predictions and forecasting. So, what I have mentioned earlier the kind of discriminant analysis, it is like linear probability model or binary choice model.

So, where l is the kind of classifiers that discriminants the between 2 different group subjects 2 different attributes. So, like this is the kind of you know, case and then similarly so, we can have here ok. So, this is how the kind of structuring, and then you can understand the particular structure, and with the basis of this particular discriminant analysis.

(Refer Slide Time: 40:45)



Classification Techniques

Classifying Credit Decisions Using Discriminant Analysis

$$L(1) = -137.48 + 32.295 \times \text{homeowner} + 0.286 \times \text{credit score} + 0.833 \times \text{years of credit history} + 0.00010274 \times \text{revolving balance} + 128.248 \times \text{revolving utilization}$$

$$L(0) = -157.2 + 30.747 \times \text{homeowner} + 0.289 \times \text{credit score} + 0.473 \times \text{years of credit history} + 0.0004716 \times \text{revolving balance} + 167.7 \times \text{revolving utilization}$$

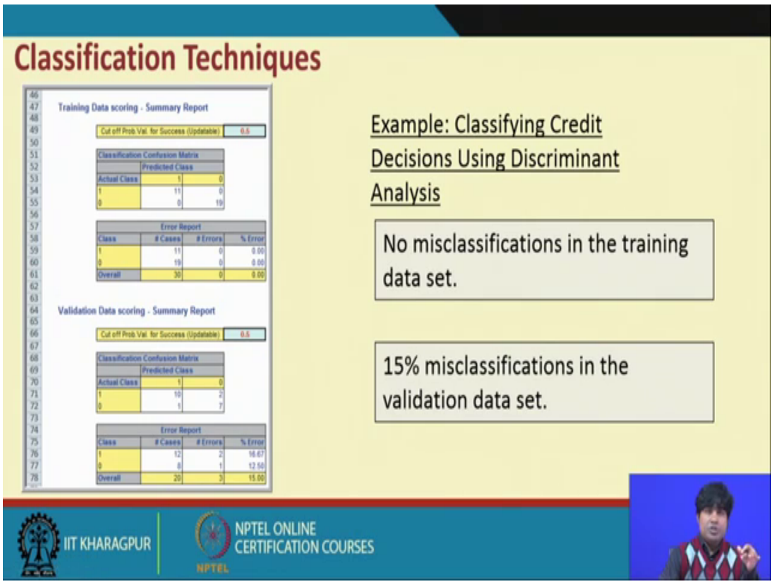
Classification Function

Variables	Classification Function	
	1	0
Constant	-137.481552	-157.201752
Homeowner	32.2950325	30.74663162
Credit Score	0.285761	0.28945312
Years of Credit History	0.83345157	0.47282016
Revolving Balance	0.00010274	0.0004716
Revolving Utilization	128.2484283	167.7003479

And the same the kind of a credit decision; which you have discussed in the case of clustering analysis so; that means, it will give you 2 different structure through which you can understand the particular requirement. And as a result, so, we have 2 different models for yes situation, and for no situation; that means, it will it will be classified into 2 different groups with respect to the particular you know, indicators.

So, this is another kind of you know, effective tools in the data mining.

(Refer Slide Time: 41:16)



Classification Techniques

Training Data scoring - Summary Report

Cut off Prob Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	11	0
0	0	19

Error Report			
Class	# Cases	# Errors	% Error
1	11	0	0.00
0	19	0	0.00
Overall	30	0	0.00

Validation Data scoring - Summary Report

Cut off Prob Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	10	2
0	1	7

Error Report			
Class	# Cases	# Errors	% Error
1	12	2	16.67
0	8	1	12.50
Overall	20	3	15.00

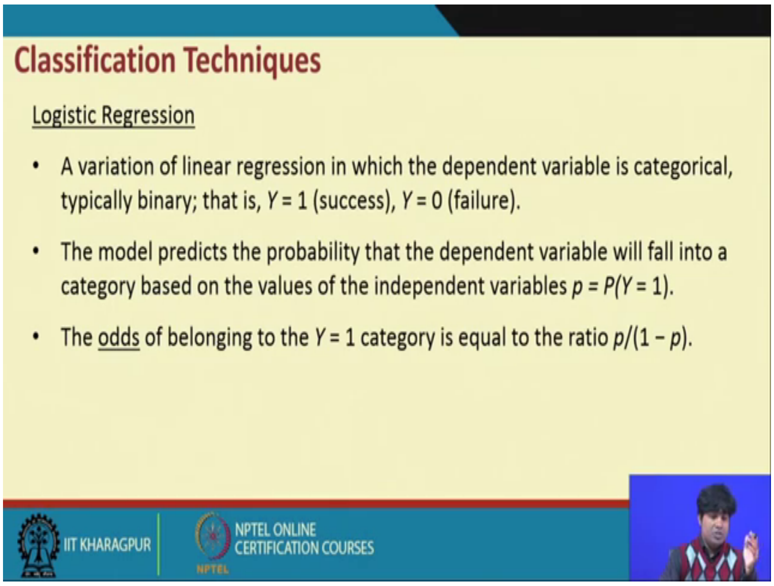
Example: Classifying Credit Decisions Using Discriminant Analysis

No misclassifications in the training data set.

15% misclassifications in the validation data set.

Through which you can understand the problem get the insights go for the kind of effective predictions, and the kind of you know, management decision qualitative response structure through which we can we have actually 3 different models to do the kind of you know, means to understand the kind of structure connect, the kind of model through, which you can do the better visualization.

(Refer Slide Time: 41:22)



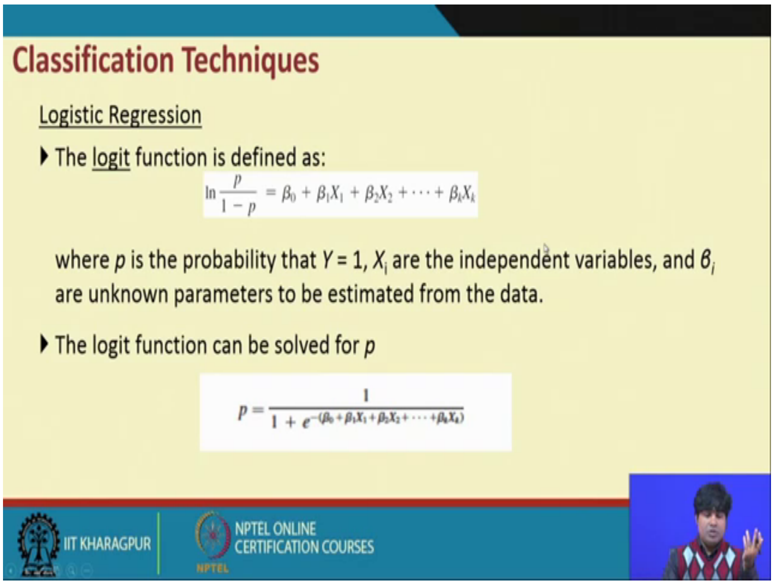
Classification Techniques

Logistic Regression

- A variation of linear regression in which the dependent variable is categorical, typically binary; that is, $Y = 1$ (success), $Y = 0$ (failure).
- The model predicts the probability that the dependent variable will fall into a category based on the values of the independent variables $p = P(Y = 1)$.
- The odds of belonging to the $Y = 1$ category is equal to the ratio $p/(1 - p)$.

Ah like the kind of clustering, then the kind of discriminant analysis. So, we have the kind of logistic structure, which we have which we have already discussed in a separate lecture. And that too in the it is a so, either you can say choose, the binary choice models or logistic models. So, or the kind of probability models ruled by you know, normal distribution function, then you come with a kind of system through which you can create a kind of set up and do the better prediction.

(Refer Slide Time: 42:05)



Classification Techniques

Logistic Regression

- The logit function is defined as:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

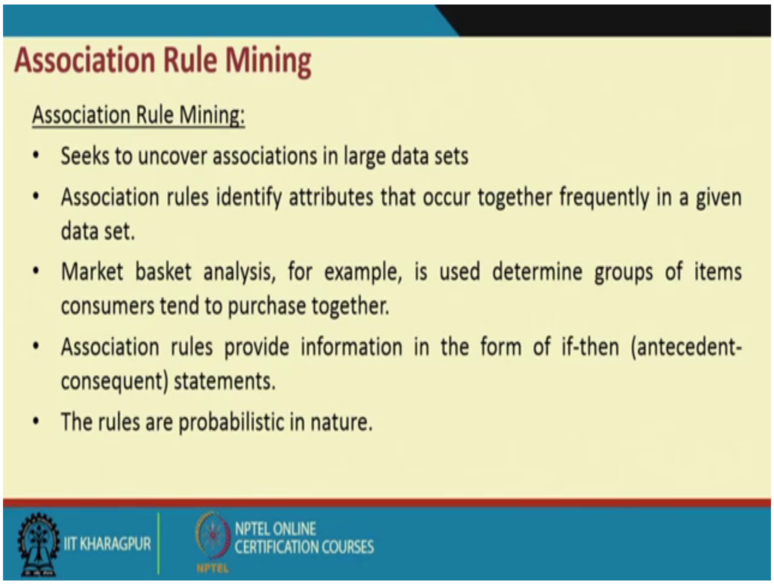
where p is the probability that $Y = 1$, X_i are the independent variables, and β_i are unknown parameters to be estimated from the data.

- The logit function can be solved for p

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

And better kind of management decisions. So, this is a simple simpler structure of logistic regression, which we have already discussed and as again. So, these are all attributes, and then the particular dependent variable structure will be classifier through which you can actually have a different kind of you know, understanding as per the particular you know, problem; and then get the insights as per the need.

(Refer Slide Time: 42:29)



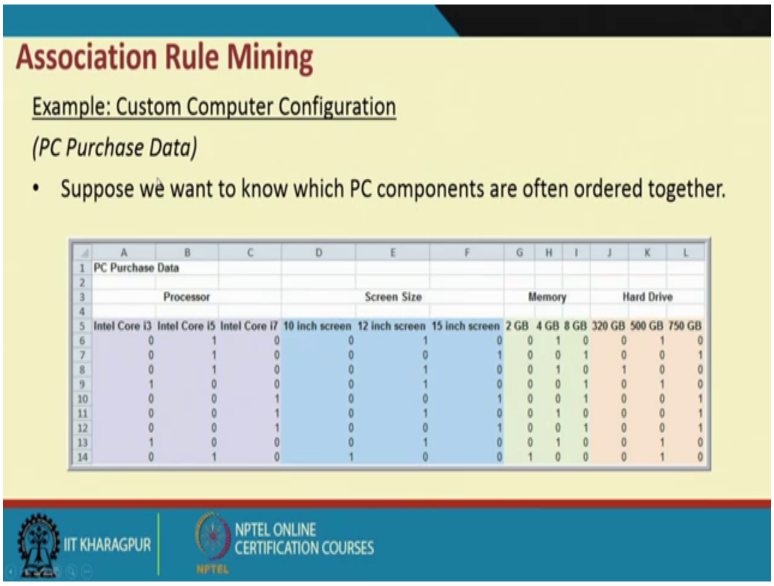
Association Rule Mining

Association Rule Mining:

- Seeks to uncover associations in large data sets
- Association rules identify attributes that occur together frequently in a given data set.
- Market basket analysis, for example, is used determine groups of items consumers tend to purchase together.
- Association rules provide information in the form of if-then (antecedent-consequent) statements.
- The rules are probabilistic in nature.

So, again so, in the data mining we have actually association structure through which you can understand the particular systems of course, we have already gone through you know, like covariance and you know, correlations, and these are the tools also can be used in the data mining to understand the particular you know, problems get the insights. And then create a structure through which you can do effective prediction, and the effective kind of you know, forecasting and that too as per the particular you know, problems need and the kind of management needs.

(Refer Slide Time: 43:05)



Association Rule Mining

Example: Custom Computer Configuration

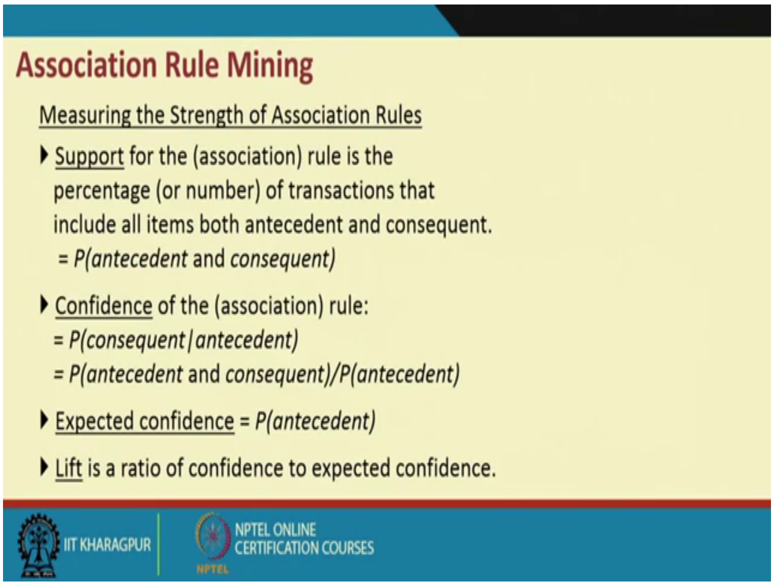
(PC Purchase Data)

- Suppose we want to know which PC components are often ordered together.

	A	B	C	D	E	F	G	H	I	J	K	L
1	PC Purchase Data											
2												
3		Processor			Screen Size			Memory			Hard Drive	
4												
5	Intel Core i3	Intel Core i5	Intel Core i7	10 inch screen	12 inch screen	15 inch screen	2 GB	4 GB	8 GB	320 GB	500 GB	750 GB
6	0	1	0	0	1	0	0	1	0	0	1	0
7	0	1	0	0	0	1	0	0	1	0	0	1
8	0	1	0	0	1	0	0	1	0	1	0	0
9	1	0	0	0	1	0	0	0	1	0	1	0
10	0	0	1	0	0	1	0	0	1	0	0	1
11	0	0	1	0	1	0	0	1	0	0	0	1
12	0	0	1	0	0	1	0	0	1	0	0	1
13	1	0	0	0	1	0	0	1	0	0	1	0
14	0	1	0	1	0	0	1	0	0	0	1	0

So; that means, we have association rules and this is a simple example through which you can actually develop a kind of you know, structures association structures; that means, we have done through clustering, we have done through classifications, again we can do analyze this problem through association technique. Then you develop a system through which you can do effective predictions and effective forecasting's.

(Refer Slide Time: 43:25)



Association Rule Mining

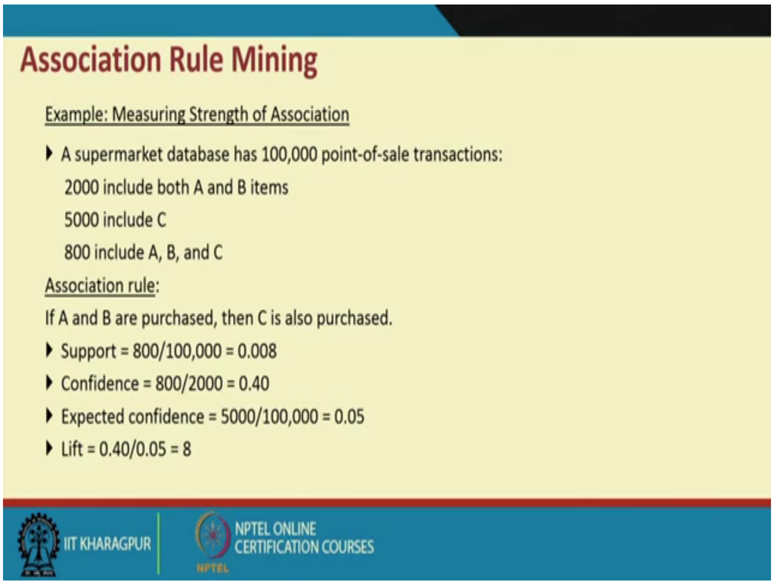
Measuring the Strength of Association Rules

- Support for the (association) rule is the percentage (or number) of transactions that include all items both antecedent and consequent.
  = *P(antecedent* and *consequent)*
- Confidence of the (association) rule:
  = *P(consequent|antecedent)*
  = *P(antecedent* and *consequent)/P(antecedent)*
- Expected confidence = *P(antecedent)*
- Lift is a ratio of confidence to expected confidence.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, this is a simple example to understand against to problems.

(Refer Slide Time: 43:30)



Association Rule Mining

Example: Measuring Strength of Association

- A supermarket database has 100,000 point-of-sale transactions:
  2000 include both A and B items
  5000 include C
  800 include A, B, and C

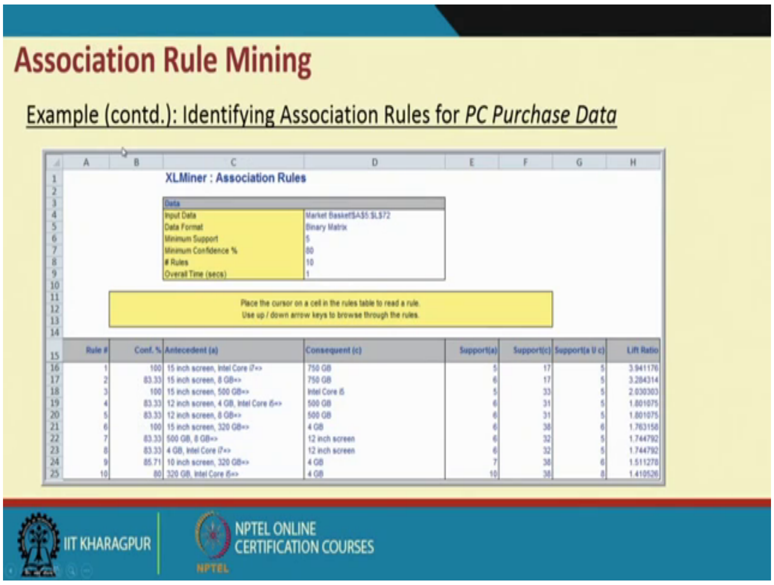
Association rule:

If A and B are purchased, then C is also purchased.

- Support = 800/100,000 = 0.008
- Confidence = 800/2000 = 0.40
- Expected confidence = 5000/100,000 = 0.05
- Lift = 0.40/0.05 = 8

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

(Refer Slide Time: 43:31)



Association Rule Mining

Example (contd.): Identifying Association Rules for *PC Purchase Data*

XLMiner : Association Rules

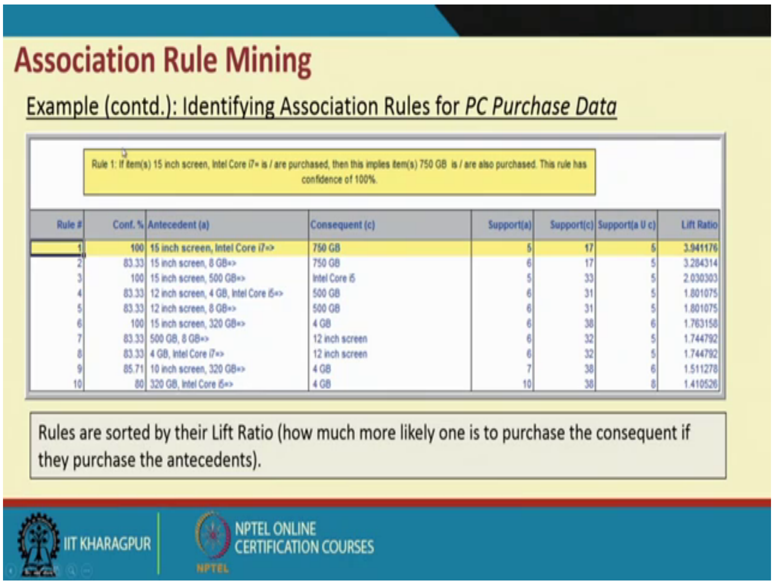
Data	
Input Data	Market Basket!$A$5:$L$72
Data Format	Binary Matrix
Minimum Support	5
Minimum Confidence %	80
# Rules	10
Overall Time (secs)	1

Place the cursor on a cell in the rules table to read a rule.
Use up / down arrow keys to browse through the rules.

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	15 inch screen, Intel Core i7=>	750 GB	5	17	5	3.941176
2	83.33	15 inch screen, 8 GB=>	750 GB	6	17	5	3.284314
3	100	15 inch screen, 500 GB=>	Intel Core i5	5	33	5	2.030303
4	83.33	12 inch screen, 4 GB, Intel Core i5=>	500 GB	6	31	5	1.801075
5	83.33	12 inch screen, 8 GB=>	500 GB	6	31	5	1.801075
6	100	15 inch screen, 320 GB=>	4 GB	6	38	6	1.763158
7	83.33	500 GB, 8 GB=>	12 inch screen	6	32	5	1.744792
8	83.33	4 GB, Intel Core i7=>	12 inch screen	6	32	5	1.744792
9	85.71	10 inch screen, 320 GB=>	4 GB	7	38	6	1.511278
10	80	320 GB, Intel Core i5=>	4 GB	10	38	8	1.410526

And these are all association structures.

(Refer Slide Time: 43:32)



Association Rule Mining

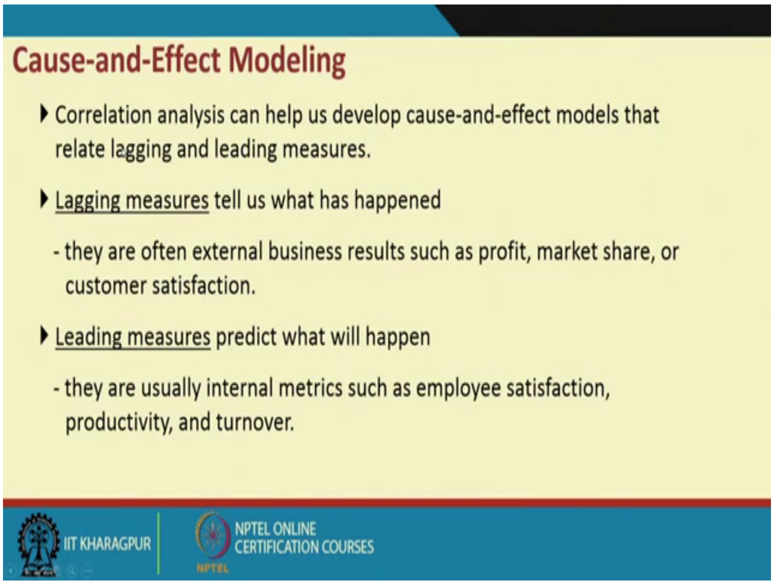
Example (contd.): Identifying Association Rules for *PC Purchase Data*

Rule 1: If item(s) 15 inch screen, Intel Core i7= is / are purchased, then this implies item(s) 750 GB is / are also purchased. This rule has confidence of 100%.

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	15 inch screen, Intel Core i7=>	750 GB	5	17	5	3.941176
2	83.33	15 inch screen, 8 GB=>	750 GB	6	17	5	3.284314
3	100	15 inch screen, 500 GB=>	Intel Core i5	5	33	5	2.030303
4	83.33	12 inch screen, 4 GB, Intel Core i5=>	500 GB	6	31	5	1.801075
5	83.33	12 inch screen, 8 GB=>	500 GB	6	31	5	1.801075
6	100	15 inch screen, 320 GB=>	4 GB	6	38	6	1.763158
7	83.33	500 GB, 8 GB=>	12 inch screen	6	32	5	1.744792
8	83.33	4 GB, Intel Core i7=>	12 inch screen	6	32	5	1.744792
9	85.71	10 inch screen, 320 GB=>	4 GB	7	38	6	1.511278
10	80	320 GB, Intel Core i5=>	4 GB	10	38	8	1.410526

Rules are sorted by their Lift Ratio (how much more likely one is to purchase the consequent if they purchase the antecedents).

(Refer Slide Time: 43:33)



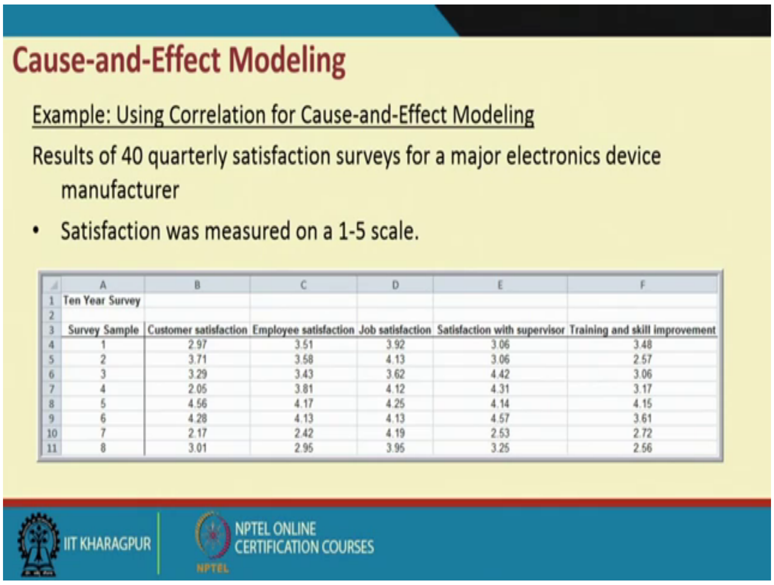
Cause-and-Effect Modeling

- Correlation analysis can help us develop cause-and-effect models that relate lagging and leading measures.
- Lagging measures tell us what has happened
  - they are often external business results such as profit, market share, or customer satisfaction.
- Leading measures predict what will happen
  - they are usually internal metrics such as employee satisfaction, productivity, and turnover.

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

And again so, there is a kind of cause and effect modeling which a in fact, the logistic structures, and binary choice models itself is a cause and effect modeling. But here also there are different steps of causality models are there through which you actually you can understand the particular you know, problem and again observe the better insights. And then think about the kind of prediction structures, and the kind of management requirement.

(Refer Slide Time: 44:07)



Cause-and-Effect Modeling

Example: Using Correlation for Cause-and-Effect Modeling

Results of 40 quarterly satisfaction surveys for a major electronics device manufacturer

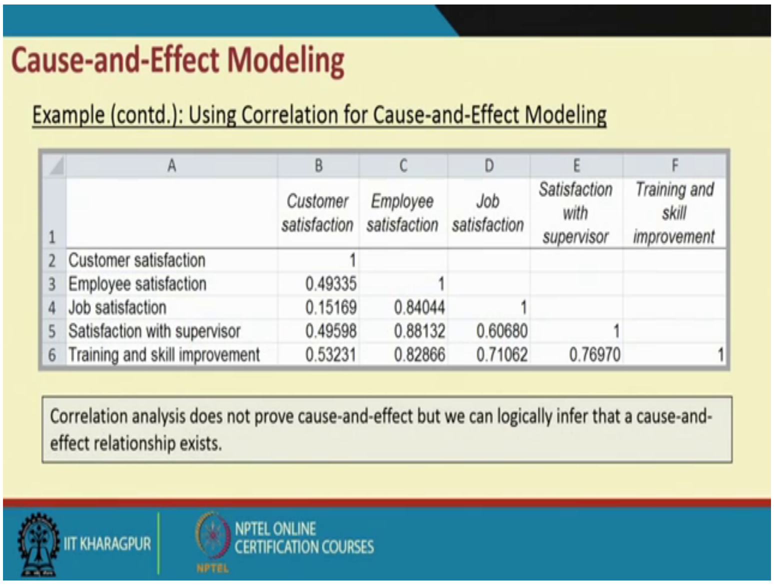
- Satisfaction was measured on a 1-5 scale.

	A	B	C	D	E	F
1	Ten Year Survey					
2						
3	Survey Sample	Customer satisfaction	Employee satisfaction	Job satisfaction	Satisfaction with supervisor	Training and skill improvement
4	1	2.97	3.51	3.92	3.06	3.48
5	2	3.71	3.58	4.13	3.06	2.57
6	3	3.29	3.43	3.62	4.42	3.06
7	4	2.05	3.81	4.12	4.31	3.17
8	5	4.56	4.17	4.25	4.14	4.15
9	6	4.28	4.13	4.13	4.57	3.61
10	7	2.17	2.42	4.19	2.53	2.72
11	8	3.01	2.95	3.95	3.25	2.56

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, that means, technically what we have mentioned we have actually plenty of you know, structure through which actually understand the problem get insights.

(Refer Slide Time: 44:09)



Cause-and-Effect Modeling

Example (contd.): Using Correlation for Cause-and-Effect Modeling

	A	B	C	D	E	F
1		Customer satisfaction	Employee satisfaction	Job satisfaction	Satisfaction with supervisor	Training and skill improvement
2	Customer satisfaction	1				
3	Employee satisfaction	0.49335	1			
4	Job satisfaction	0.15169	0.84044	1		
5	Satisfaction with supervisor	0.49598	0.88132	0.60680	1	
6	Training and skill improvement	0.53231	0.82866	0.71062	0.76970	1

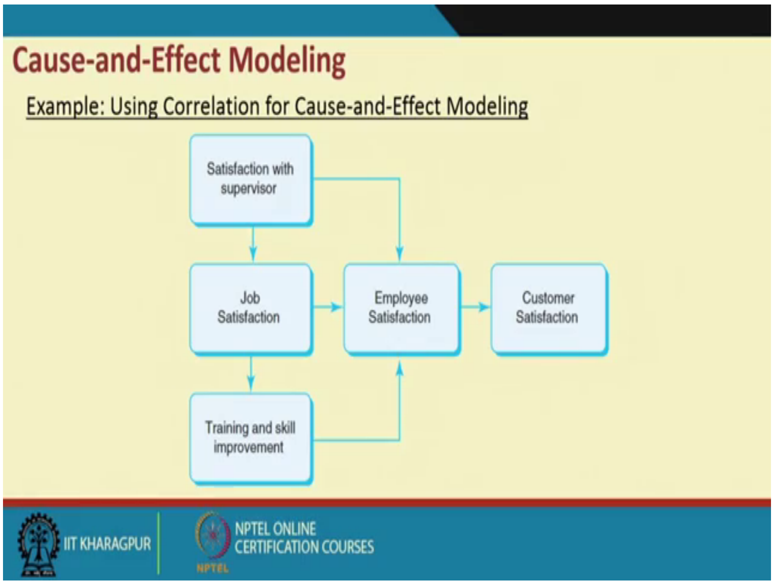
Correlation analysis does not prove cause-and-effect but we can logically infer that a cause-and-effect relationship exists.

Then you know, think about the final set up through which you can do the effective structure, and do the kind of prediction and forecasting as per the particular you know, business requirement and the kind of management requirement.

So, that means, actually data mining is a kind of systems, we have actually n number of you know, tools are there analytics tools are there starting with you know, simple classification clustering. Then the kind of association cause and effect models and depending upon a particular problem structures the kind of data structures then the kind of understanding, the kind of business objectives or management objectives.

So, you pick up a particular tools and then analyze the particular problems, then get the better insights and then develop the system exactly as per the particular you know, business requirement and management for instance in the cause and effect model.

(Refer Slide Time: 45:16)



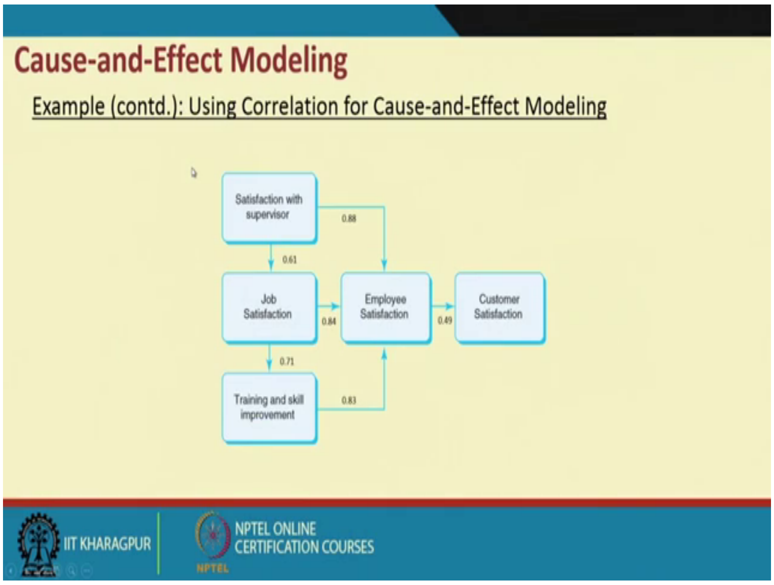
Cause-and-Effect Modeling

Example: Using Correlation for Cause-and-Effect Modeling

Satisfaction with supervisor
Job Satisfaction
Employee Satisfaction
Customer Satisfaction
Training and skill improvement
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

And this is how the simple example and where the customer satisfaction is the kind of you know, major effect. And then this is derived through job satisfaction training and skill improvement satisfaction with supervisor employee satisfaction. So, that means, in this particular case the most important thing is how to develop the particular you know, structure and then data will help you to verify the particular you know, structure and then we can actually validate and then use this particular model for the prediction. And you know, forecasting or as per the particular you know, management requirement again with the data.

(Refer Slide Time: 45:54)



Cause-and-Effect Modeling

Example (contd.): Using Correlation for Cause-and-Effect Modeling

So, we can have here the kind of you know, outcome and these are all correlation output through which you know, because we are discussing the cause and effect and association through which you can do the kind of better prediction. Now with the data, what we have observed here the correlation coefficients are very high as a result. So, the particular model's you know, for this particular problem is very effective for the kind of you know, management requirement.

So, likewise we can actually think about the particular problems, check the data structure you know, pick up a particular technique, and then develop the system, which can be very effective as per the particular you know, problem requirement or you know, management requirement. With this we will stop here.

Thank you very much; have a nice day.