PySpark Certification T...

Deep Dive into Spark MLlib > Presentation

Course Content

- Presentation
- Class 8 Recording
- Case Study I - Use di...
- Case Study II - Use ...
- In-Class Demo - Inpu...
- Dataset
- Dataset
- Spam Detection Code
- Understanding Apache...

Personal Library

Presentation 3 / 139

COURSE OUTLINE

MODULE 08

Introduction to Big Data Hadoop and Spark

Introduction to Python for Apache Spark

Functions, OOPs, Modules in Python

Deep Dive into Apache Spark Framework

Playing with Spark RDDs

Data Frames and Spark SQL

Machine Learning using Spark MLlib

Deep Dive into Spark MLlib

Understanding Apache Kafka and Apache Flume

Apache Spark Streaming – Processing Multiple Batches

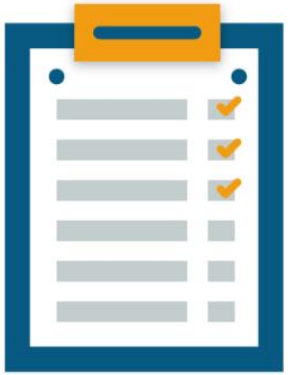
Apache Spark Streaming - Data Sources

Implementing an End-to-End Project

Presentation 4 / 139

Topics

- Supervised Learning
- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes



edureka!

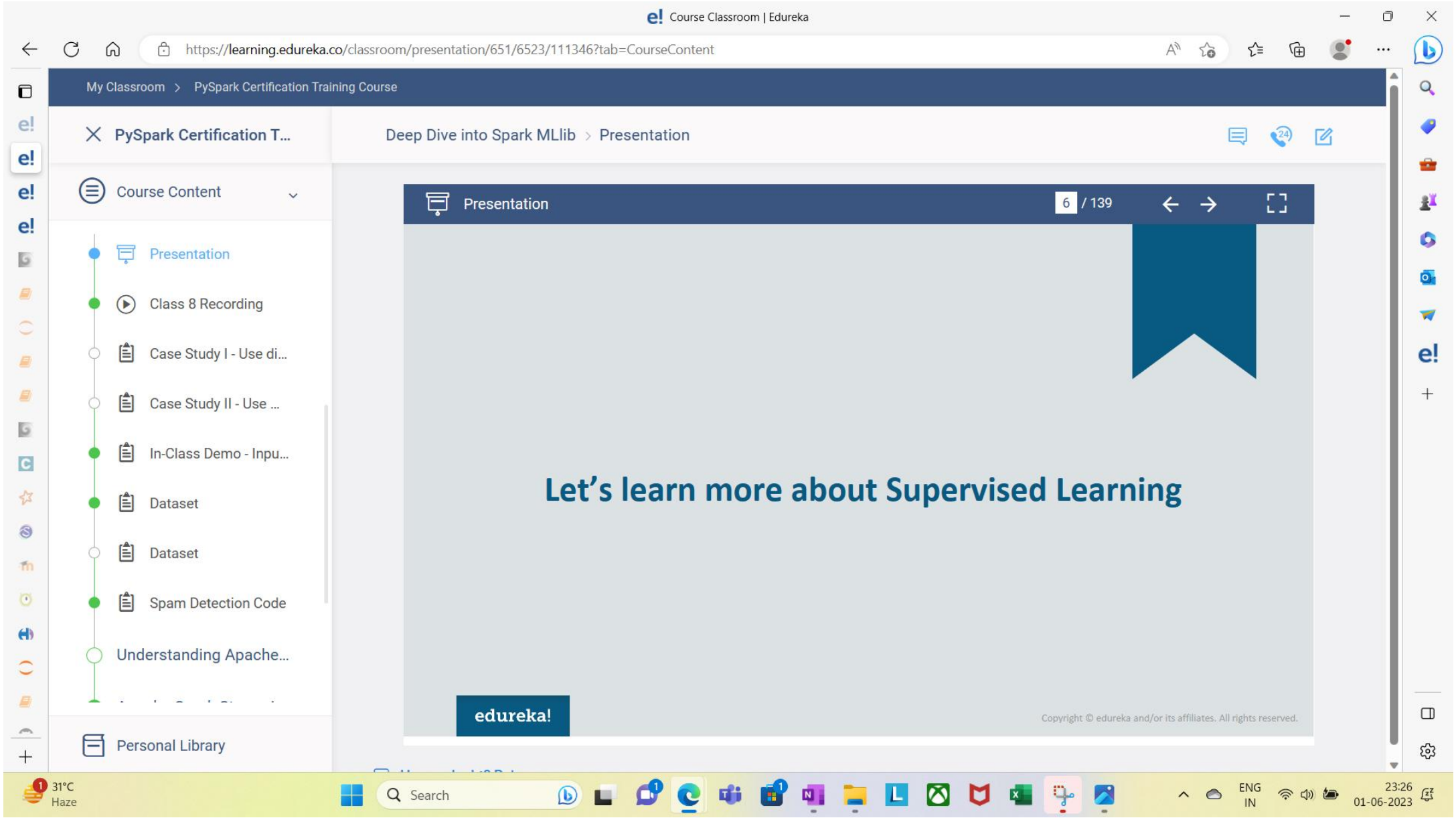
Presentation 5 / 139

Objectives

After completing this module, you should be able to:

- Illustrate Supervised Learning Algorithms
  - Linear Regression
  - Logistic Regression
  - Decision Tree
  - Random Forest
- Define Unsupervised Learning
- Discuss the Cluster Analysis
  - K-Means Clustering

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
6 / 139
Let's learn more about Supervised Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

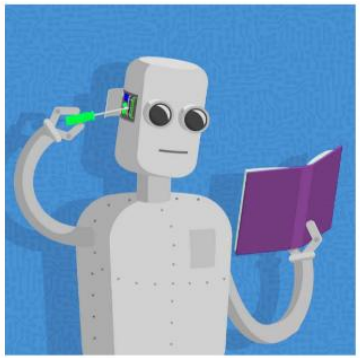
Presentation 7 / 139

Supervised Learning

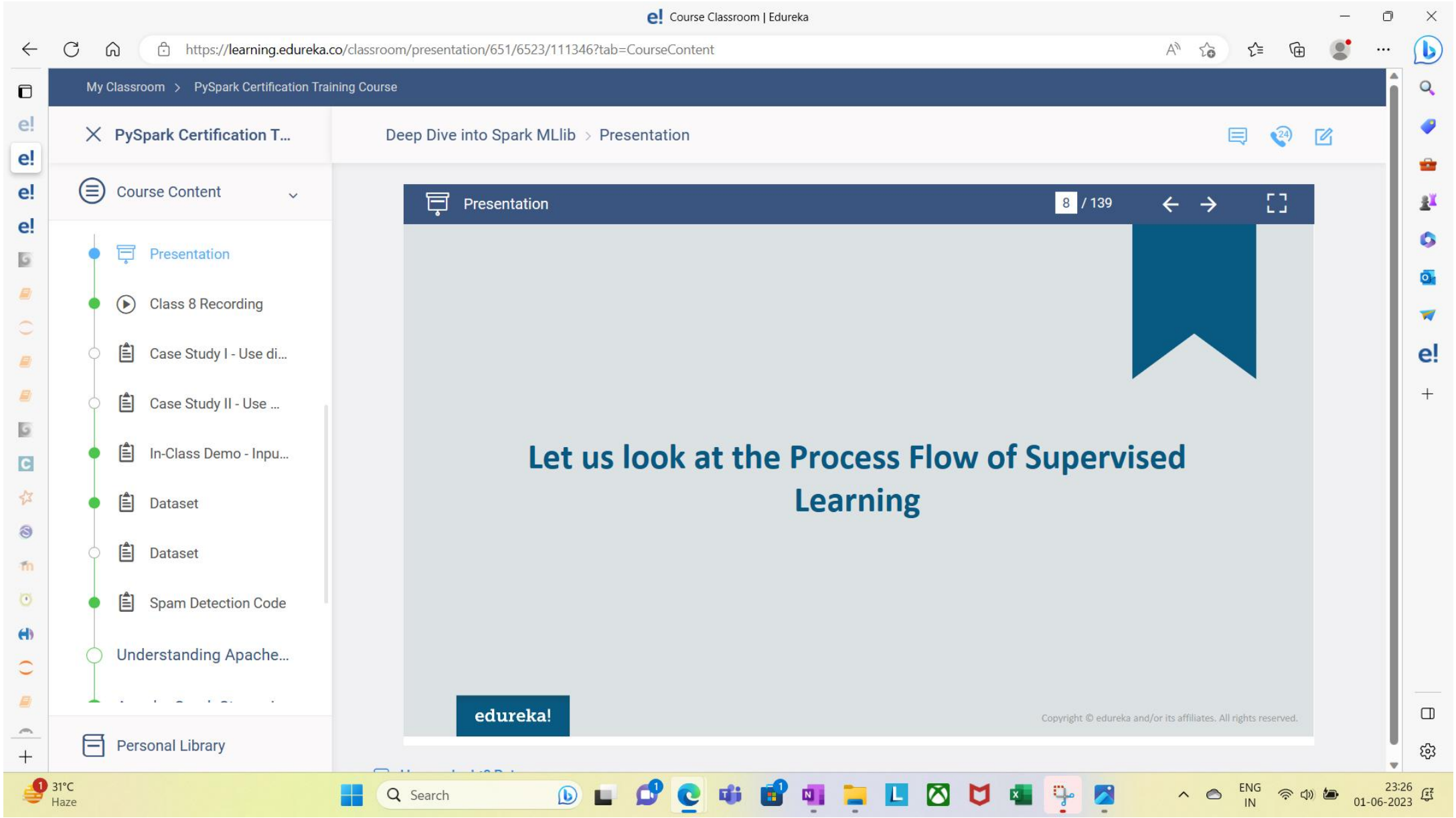
Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output

Function: Y=F(X)

It is called Supervised Learning because the process of an algorithm learning from the training dataset can be thought as a teacher supervising the learning process

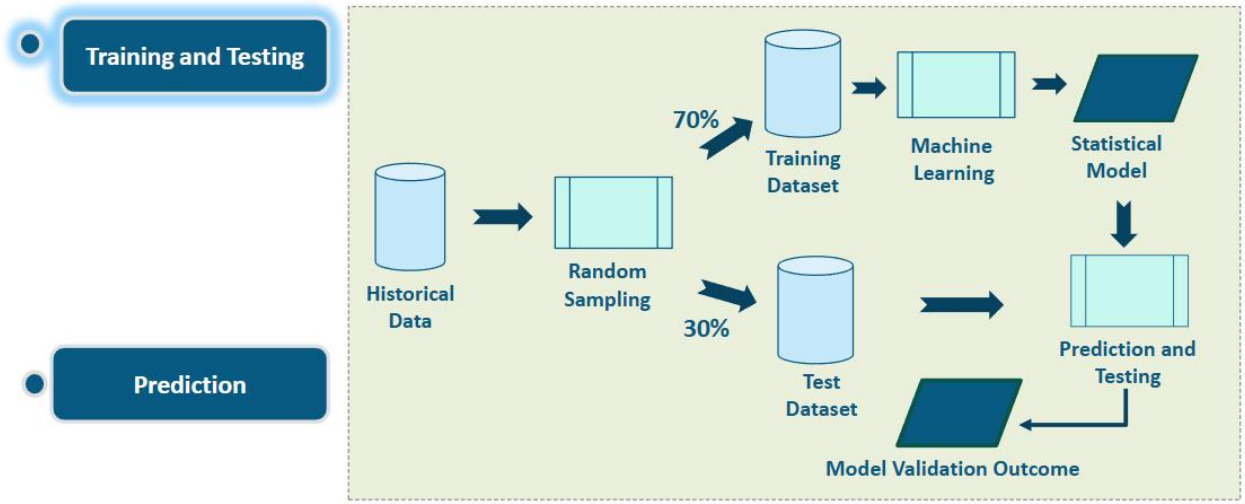


edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
8 / 139
Let us look at the Process Flow of Supervised Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

Presentation 9 / 139

Process Flow: Supervised Learning



edureka!

Process Flow: Supervised Learning

Training and Testing

Prediction

The model is used for predicting outcome of a new data set. Whenever performance of the model degrades, the model is retrained

edureka!

Presentation 11 / 139

Supervised Learning Algorithms

Used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s)

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes Classifier

edureka!

Presentation 12 / 139

Supervised Learning Algorithms

- Linear Regression — Used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s)
- Logistic Regression — Used to estimate discrete values (binary values like 0/1, yes/no, true/false) based on given set of independent variable(s)
- Decision Tree
- Random Forest
- Naïve Bayes Classifier

edureka!

Presentation 13 / 139

Supervised Learning Algorithms

- Linear Regression — Used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s)
- Logistic Regression — Used to estimate discrete values (binary values like 0/1, yes/no, true/false) based on given set of independent variable(s)
- Decision Tree — Used for classification problems. It works for both categorical and continuous dependent variables
- Random Forest
- Naïve Bayes Classifier

edureka!

Presentation 14 / 139

Supervised Learning Algorithms

Linear Regression	Used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s)
Logistic Regression	Used to estimate discrete values (binary values like 0/1, yes/no, true/false) based on given set of independent variable(s)
Decision Tree	Used for classification problems. It works for both categorical and continuous dependent variables
Random Forest	Random Forest is an ensemble of decision trees. It gives better prediction and accuracy than decision tree
Naïve Bayes Classifier	

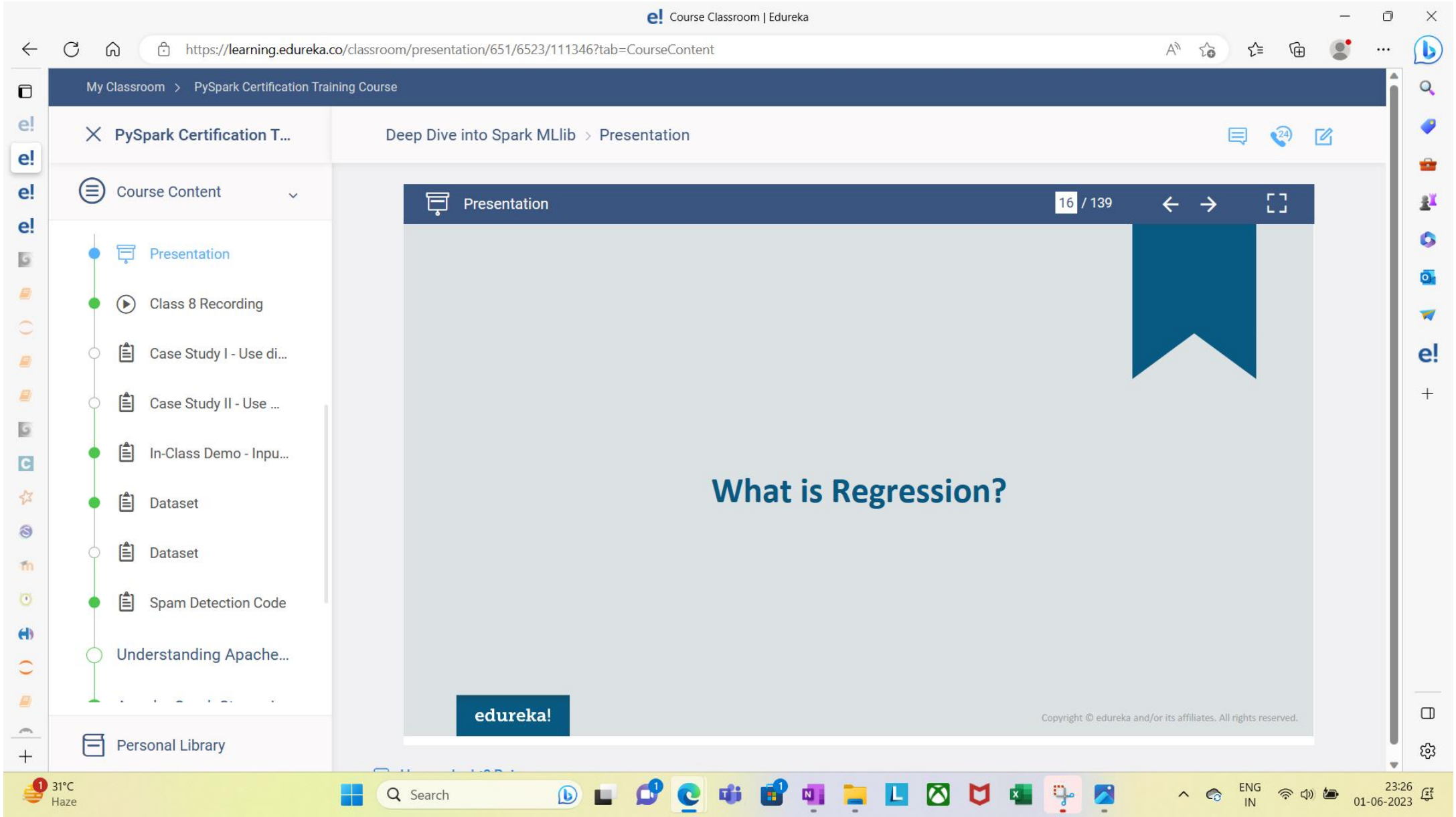
edureka!

Presentation 15 / 139

Supervised Learning Algorithms

Linear Regression	Used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s)
Logistic Regression	Used to estimate discrete values (binary values like 0/1, yes/no, true/false) based on given set of independent variable(s)
Decision Tree	Used for classification problems. It works for both categorical and continuous dependent variables
Random Forest	Random Forest is an ensemble of decision trees. It gives better prediction and accuracy than decision tree
Naïve Bayes Classifier	It is a classification technique based on Bayes' theorem with an assumption of independence between predictors

edureka!

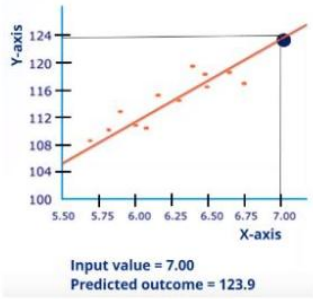

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
16 / 139
What is Regression?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

Presentation 17 / 139

What is Regression?

Regression Analysis is a predictive modelling technique

It estimates the relationship between a dependent (target) and an independent variable(predictor)

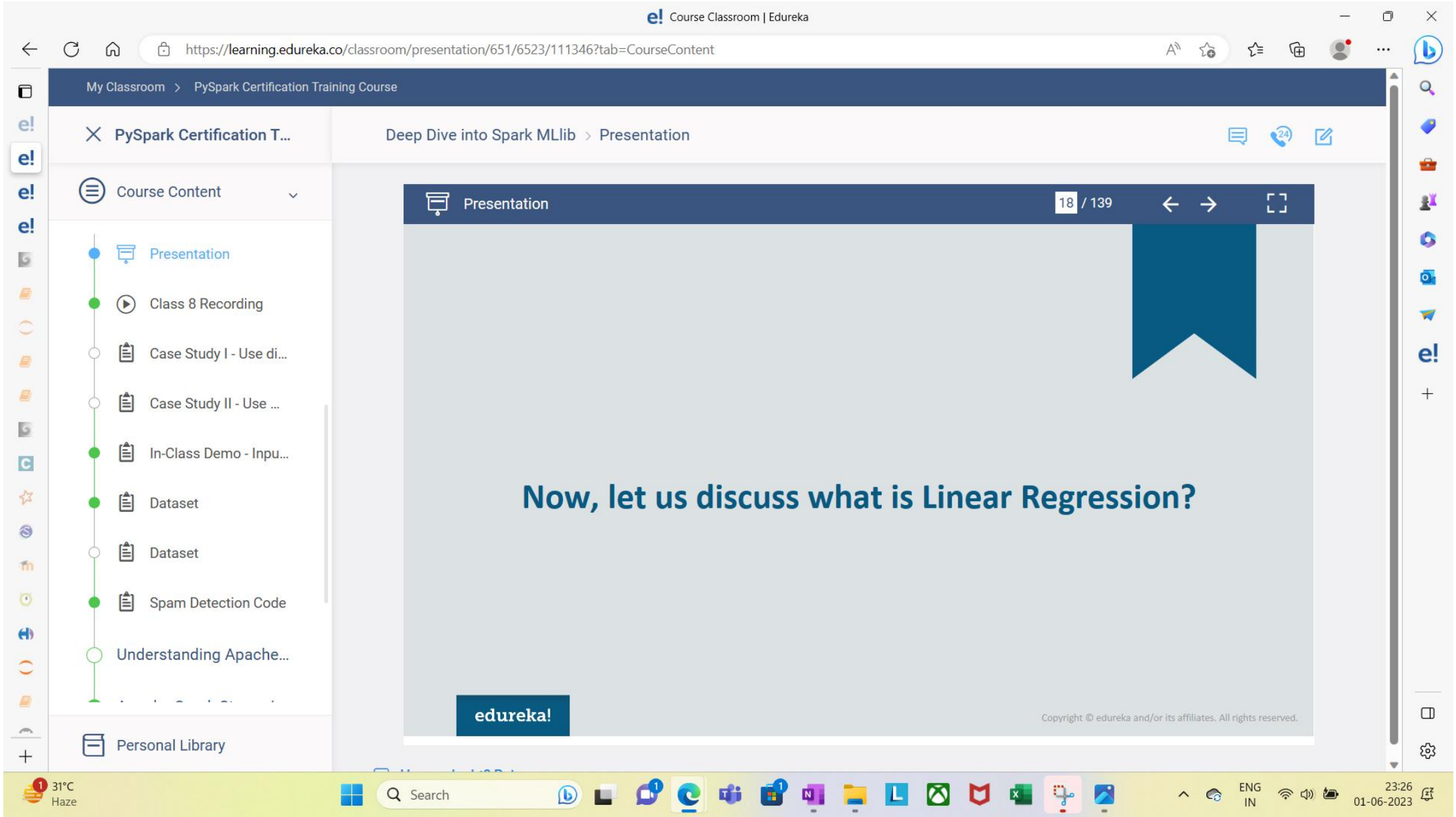


In this graph, Y-variable is dependent on X-variable

X-Variable can increase it's value as much as it wants. So, for an arbitrary value of x, we are "predicting" value of Y

This prediction is done using Regression

edureka!

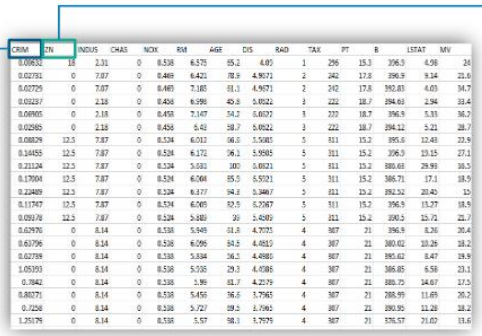

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
18 / 139
Now, let us discuss what is Linear Regression?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

Presentation 19 / 139

What is Linear Regression?

- Linear Regression Analysis is a powerful technique used for predicting the unknown value of a variable (Dependent Variable) from the known value of another variables (Independent Variable)
- A **Dependent Variable(DV)** is the variable to be predicted or explained in a regression model
- An **Independent Variable(IDV)** is the variable related to the dependent variable in a regression equation

For Example:-



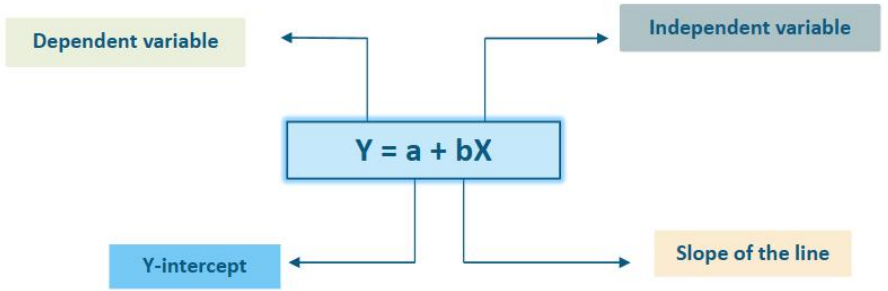
Independent Variable

Dependent Variable

edureka!

Presentation 20 / 139

Simple Linear Regression



- **Y-intercept (a)** is that value of the Dependent Variable(y) when the value of the Independent Variable(x) is zero. It is the point at which the line cuts the y-axis.
- **Slope (b)** is the change in the Dependent Variable for a unit increase in the Independent Variable. It is the tangent of the angle made by the line with the x-axis.

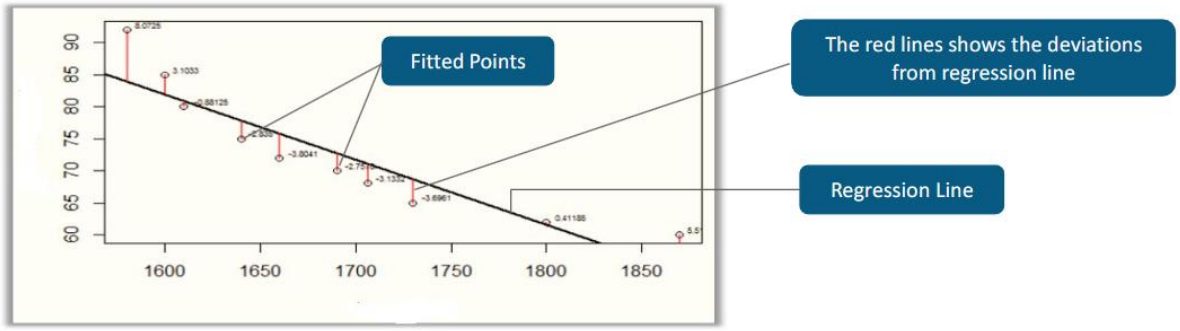
edureka!

Presentation 21 / 139

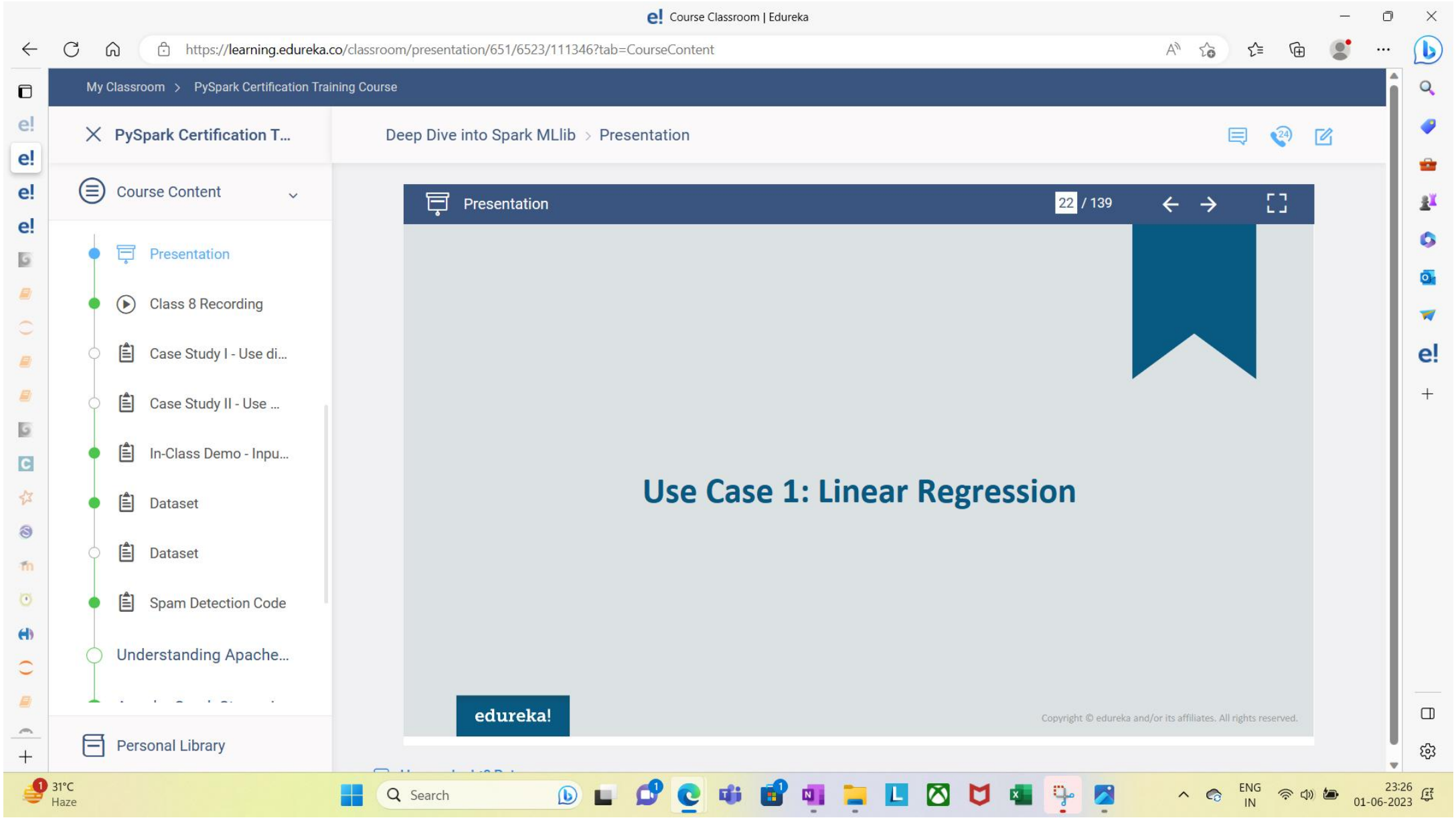
The Regression Line

The regression line is simply a single line that best fits the data (in terms of having the smallest overall distance from the line to the points)

This technique is used for finding the "best-fitting line" using the "least squares method".



edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
22 / 139
Use Case 1: Linear Regression
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

Presentation 23 / 139

Use Case 1: Linear Regression

For Example:-

Political elections are being contested in our country

Suppose that we are interested to know which candidate will probably win

The Outcome variable(result) is binary, either win or lose

The predictor variables are amount of money spent, age, popularity rank etc

Output

election_id	result	year	amount_spent	popularity_rank
122	0	32	3.81	3
315	1	48	6.32	2
201	1	51	3.67	1
965	0	40	2.93	4
410	1	52	3.6	1
150	0	35	4.2	4
743	1	39	5.66	2
612	1	42	4.32	3
206	1	44	3.26	3
792	0	50	4.52	4

Historical Data

edureka!

Presentation 24 / 139

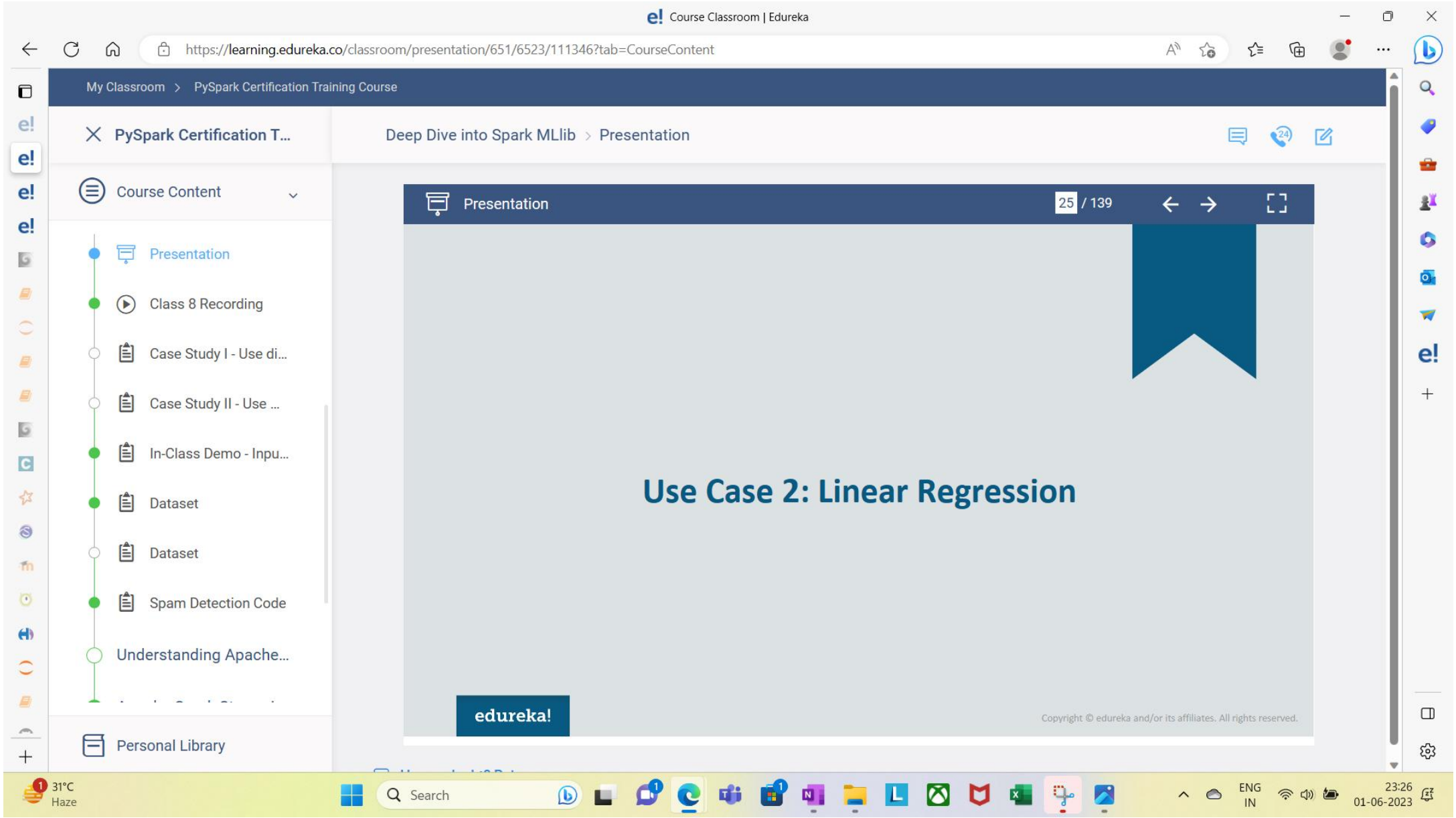
Use Case 1: Linear Regression

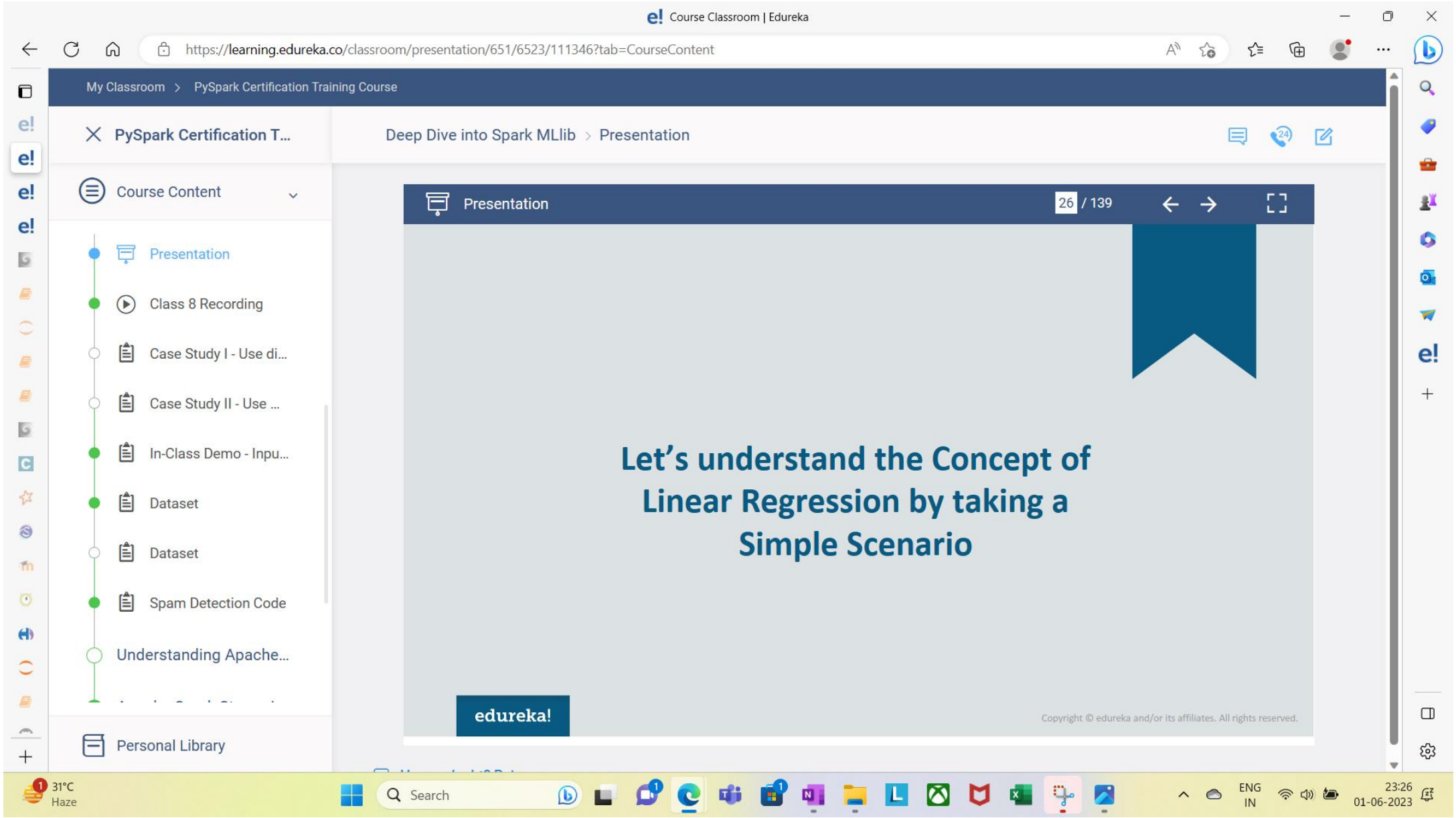
For Example:-

Political elections are being contested in our country

Suppose that we are interested to know which candidate will probably win

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
25 / 139
Use Case 2: Linear Regression
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
26 / 139
Let's understand the Concept of Linear Regression by taking a Simple Scenario
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

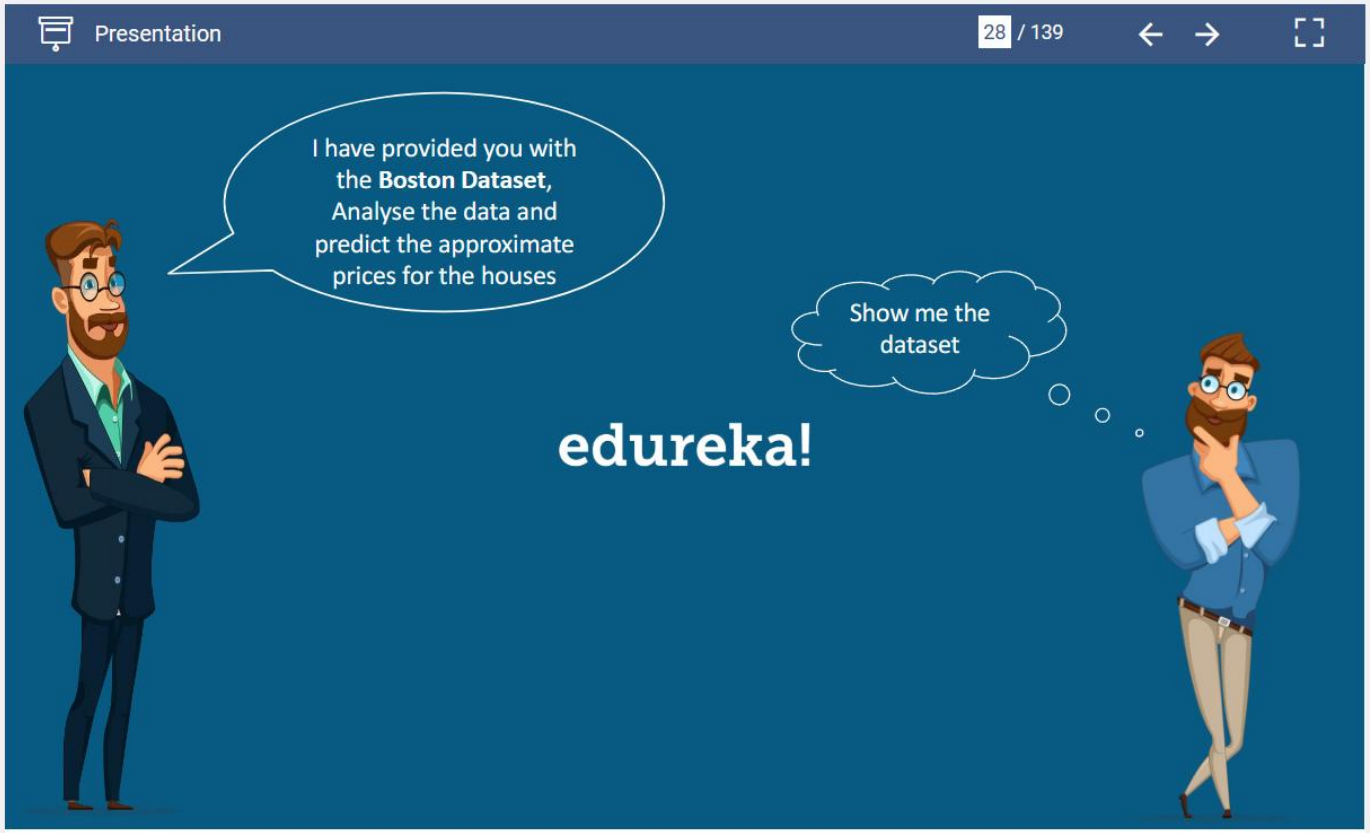
Presentation 27 / 139

Use Case 2: Linear Regression – A Real Estate Company

A Real Estate Company "Prime Homes" has a new project coming up in which they have build the homes at different locations in Boston.

They have rough idea about prices but actual price is not decided yet. They want prices such that houses can be easily afforded by common people.

edureka!



Boston Dataset

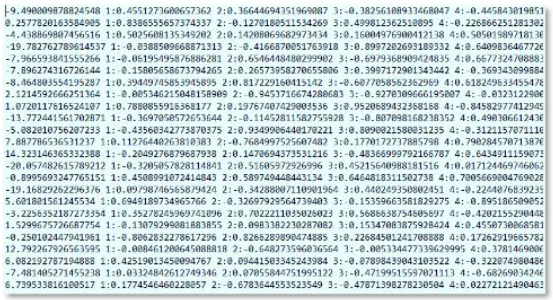
Here is the Boston data which "Prime Homes" will use to predict the price:-

	0	1	2	3	4
summary	count	mean	stddev	min	max
CRIM	506	3.6135235608162057	8.601545086715594	0.00632	88.97619629
ZN	506	11.363636363636363	23.32245299451514	0.0	100.0
INDUS	506	11.136778749531626	6.86035298095724	0.460000008	27.73999977
CHAS	506	0.0691699604743083	0.2539940413404101	0	1
NOX	506	0.5546950602312246	0.1158776754570543	0.38499999	0.870999992
RM	506	6.28463438896641	0.7026171549511354	3.561000109	8.779999733
AGE	506	68.57490120115612	28.148861532793276	2.900000095	100.0
DIS	506	3.7950426960059325	2.105710142043288	1.129600048	12.12650013
RAD	506	9.549407114624506	8.707259384239366	1	24
TAX	506	408.2371541501976	168.53711605495903	187	711
PT	506	18.45553382776679	2.164945780039869	12.60000038	22.0
B	506	356.67402960597883	91.29486340272308	0.319999993	396.8999939
LSTAT	506	12.653063233922925	7.141061500195388	1.730000019	37.97000122
MV	506	22.53280636250988	9.197104107945272	5.0	50.0

edureka!

Boston Dataset Description

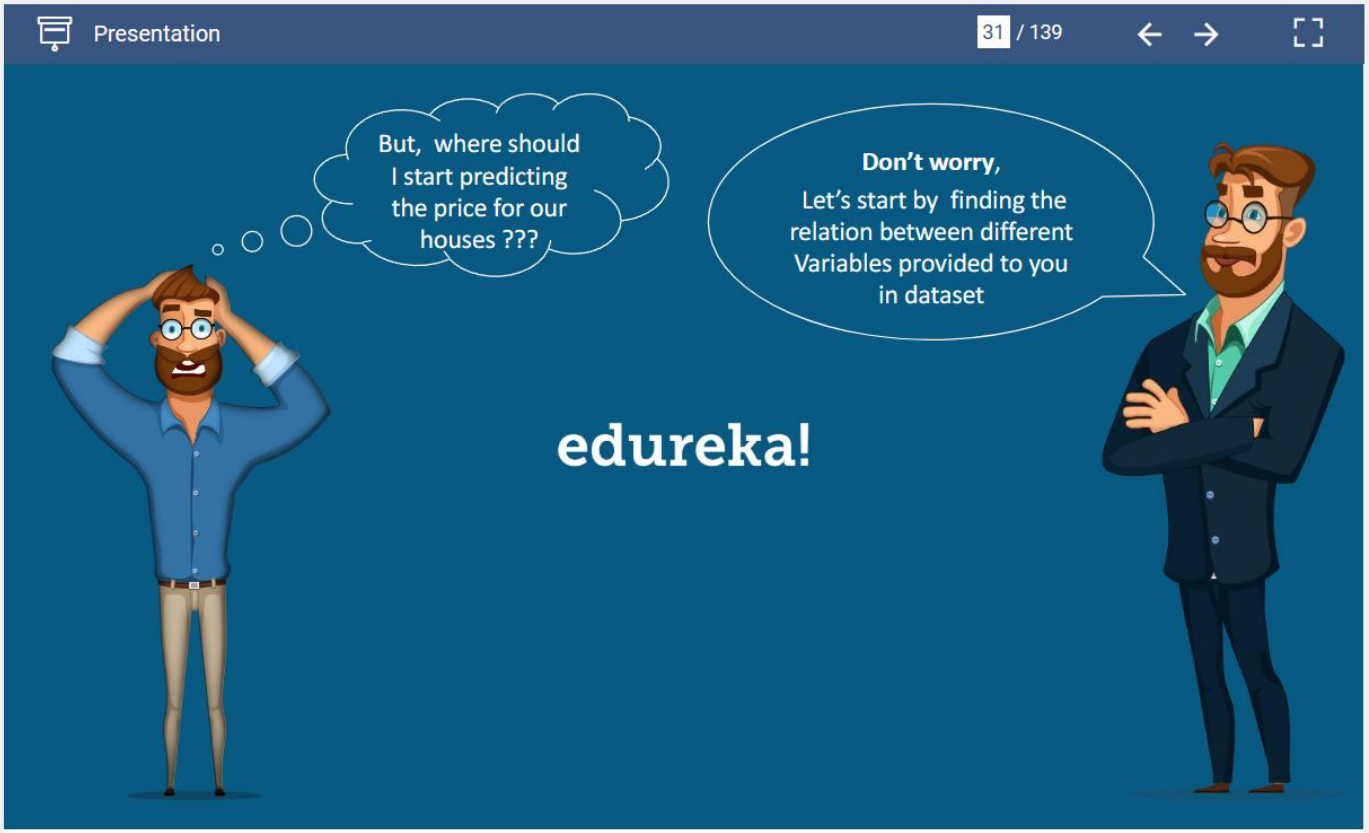
In order to train the model, we will use Boston dataset. The dataset looks like this:

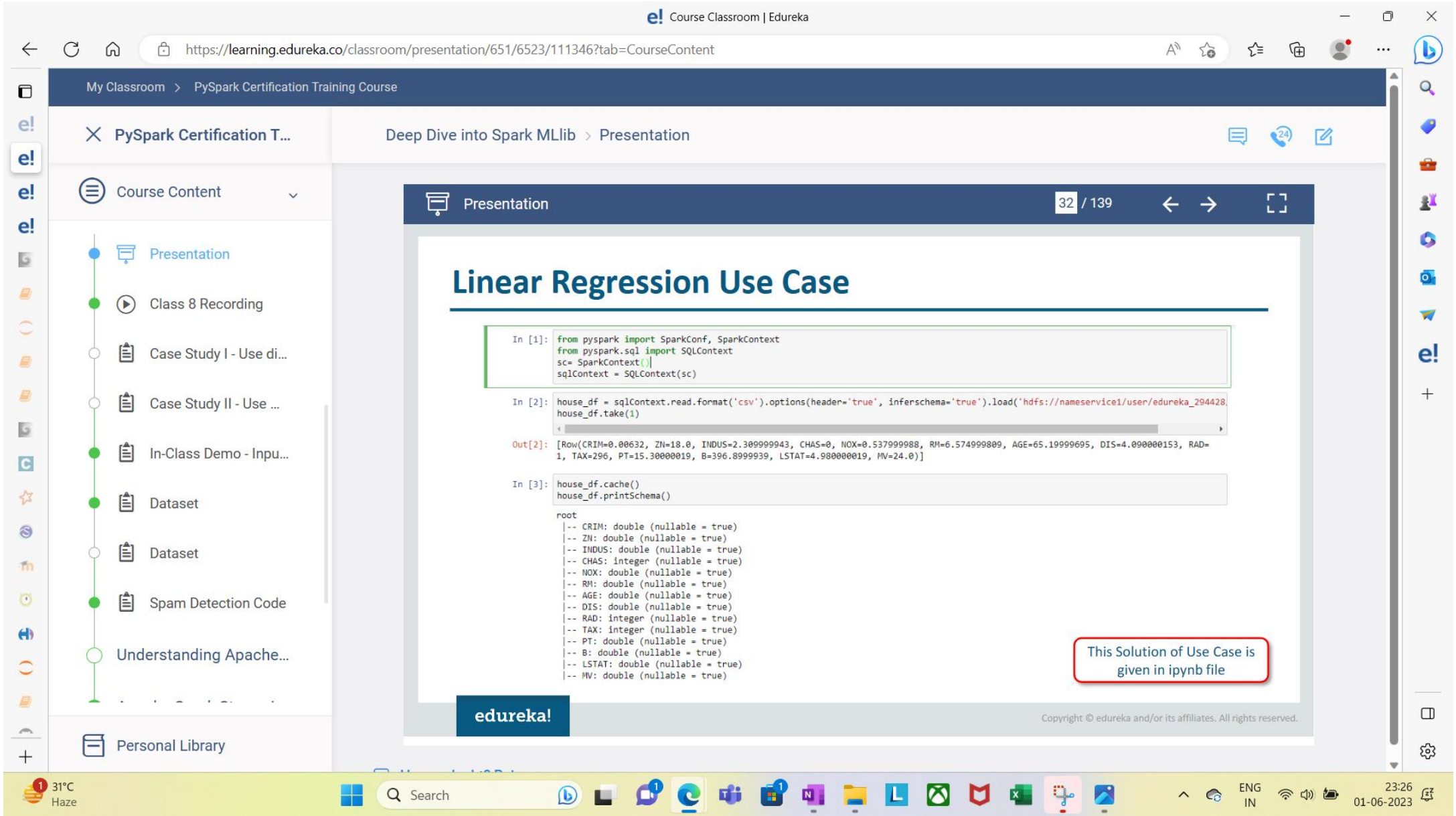


- *CRIM* - per capita crime rate by town
- *ZN* - proportion of residential land zoned for lots over 25,000 sq.ft.
- *INDUS* - proportion of non-retail business acres per town.
- *CHAS* - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- *NOX* - nitric oxides concentration (parts per 10 million)
- *RM* - average number of rooms per dwelling
- *AGE* - proportion of owner-occupied units built prior to 1940
- *DIS* - weighted distances to five Boston employment centres
- *RAD* - index of accessibility to radial highways
- *TAX* - full-value property-tax rate per $10,000
- *PT* - pupil-teacher ratio by town
- *B* - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- *LSTAT* - % lower status of the population
- *MV* - Median value of owner-occupied homes in $1000's

edureka!

Presentation 31 / 139

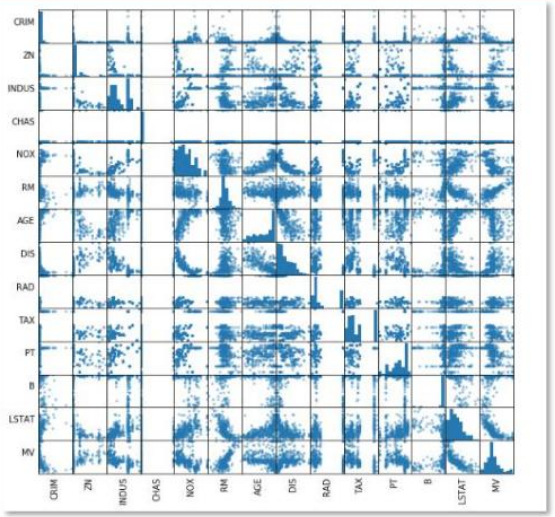



Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
32 / 139
Linear Regression Use Case
In [1]:
from pyspark import SparkConf, SparkContext
from pyspark.sql import SQLContext
sc= SparkContext()
sqlContext = SQLContext(sc)
In [2]:
house_df = sqlContext.read.format('csv').options(header='true', inferschema='true').load('hdfs://nameservice1/user/edureka_294428
house_df.take(1)
Out[2]:
[Row(CRIM=0.00632, ZN=18.0, INDUS=2.309999943, CHAS=0, NOX=0.537999988, RM=6.574999809, AGE=65.19999695, DIS=4.090000153, RAD=
1, TAX=296, PT=15.30000019, B=396.8999939, LSTAT=4.980000019, MV=24.0)]
In [3]:
house_df.cache()
house_df.printSchema()
root
|-- CRIM: double (nullable = true)
|-- ZN: double (nullable = true)
|-- INDUS: double (nullable = true)
|-- CHAS: integer (nullable = true)
|-- NOX: double (nullable = true)
|-- RM: double (nullable = true)
|-- AGE: double (nullable = true)
|-- DIS: double (nullable = true)
|-- RAD: integer (nullable = true)
|-- TAX: integer (nullable = true)
|-- PT: double (nullable = true)
|-- B: double (nullable = true)
|-- LSTAT: double (nullable = true)
|-- MV: double (nullable = true)
This Solution of Use Case is given in ipynb file
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C Haze
Search
ENG IN
23:26 01-06-2023

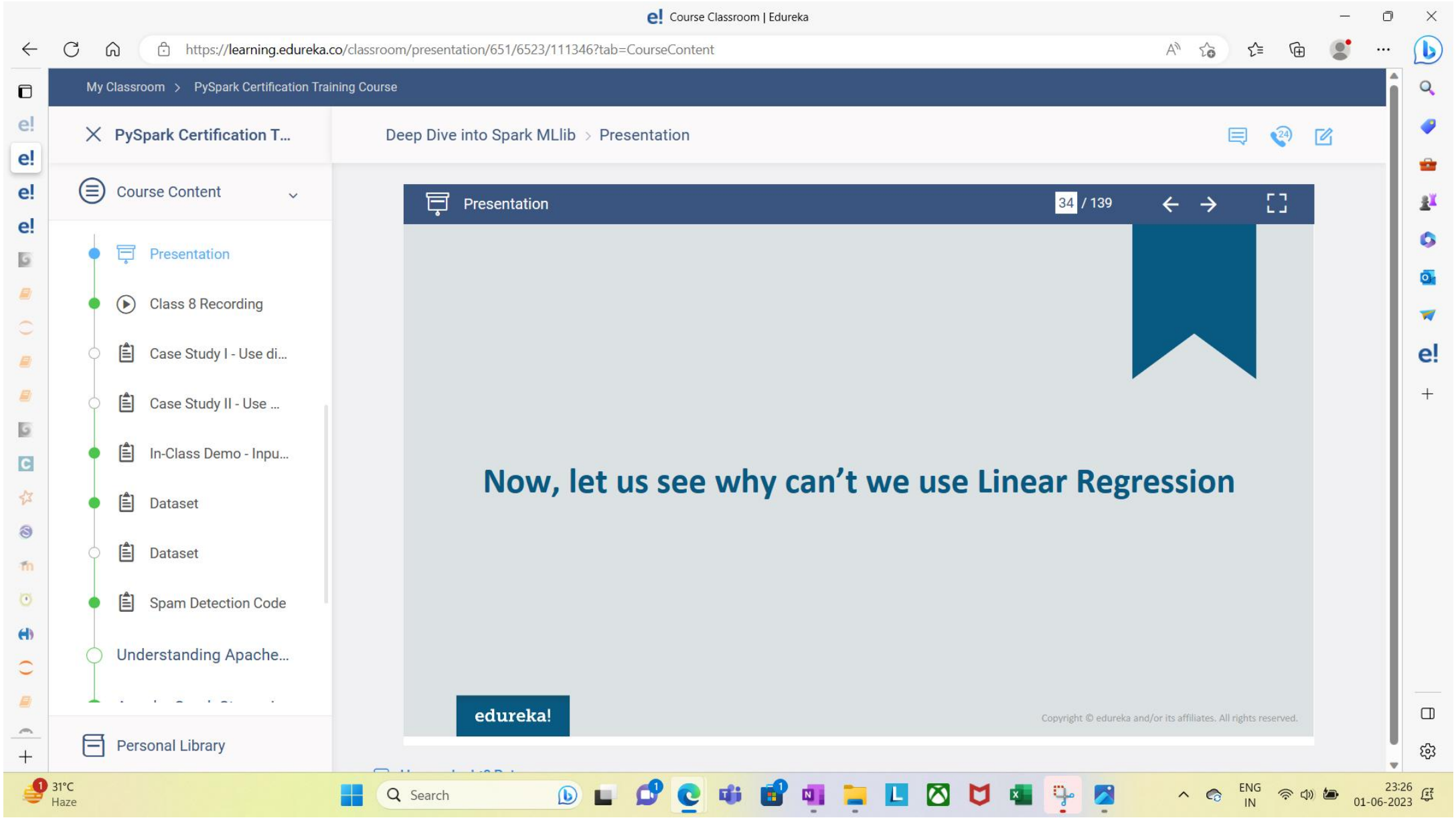
Presentation 33 / 139

Relation Between Dependent and Independent Variable

In order to know how these variables are related to each other, we plot them against each other

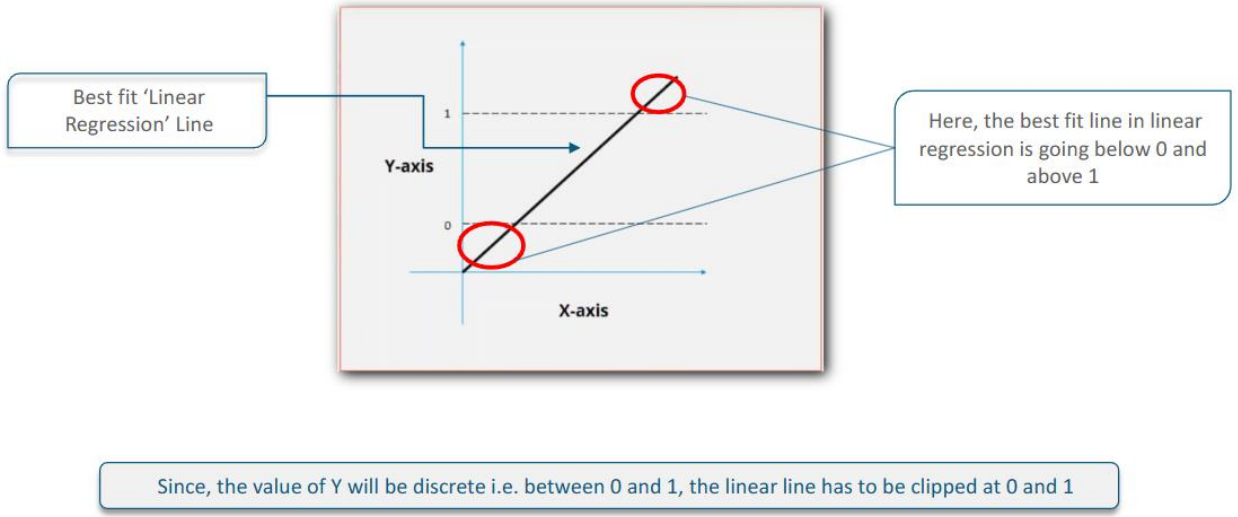


edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
34 / 139
Now, let us see why can't we use Linear Regression
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

Presentation 35 / 139

Why Not Linear Regression?



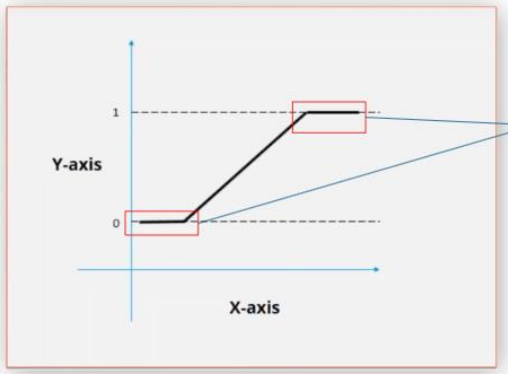
Since, the value of Y will be discrete i.e. between 0 and 1, the linear line has to be clipped at 0 and 1

edureka!

Presentation 36 / 139

Why Not Linear Regression?

Linear regression gives us one single line. To classify the output



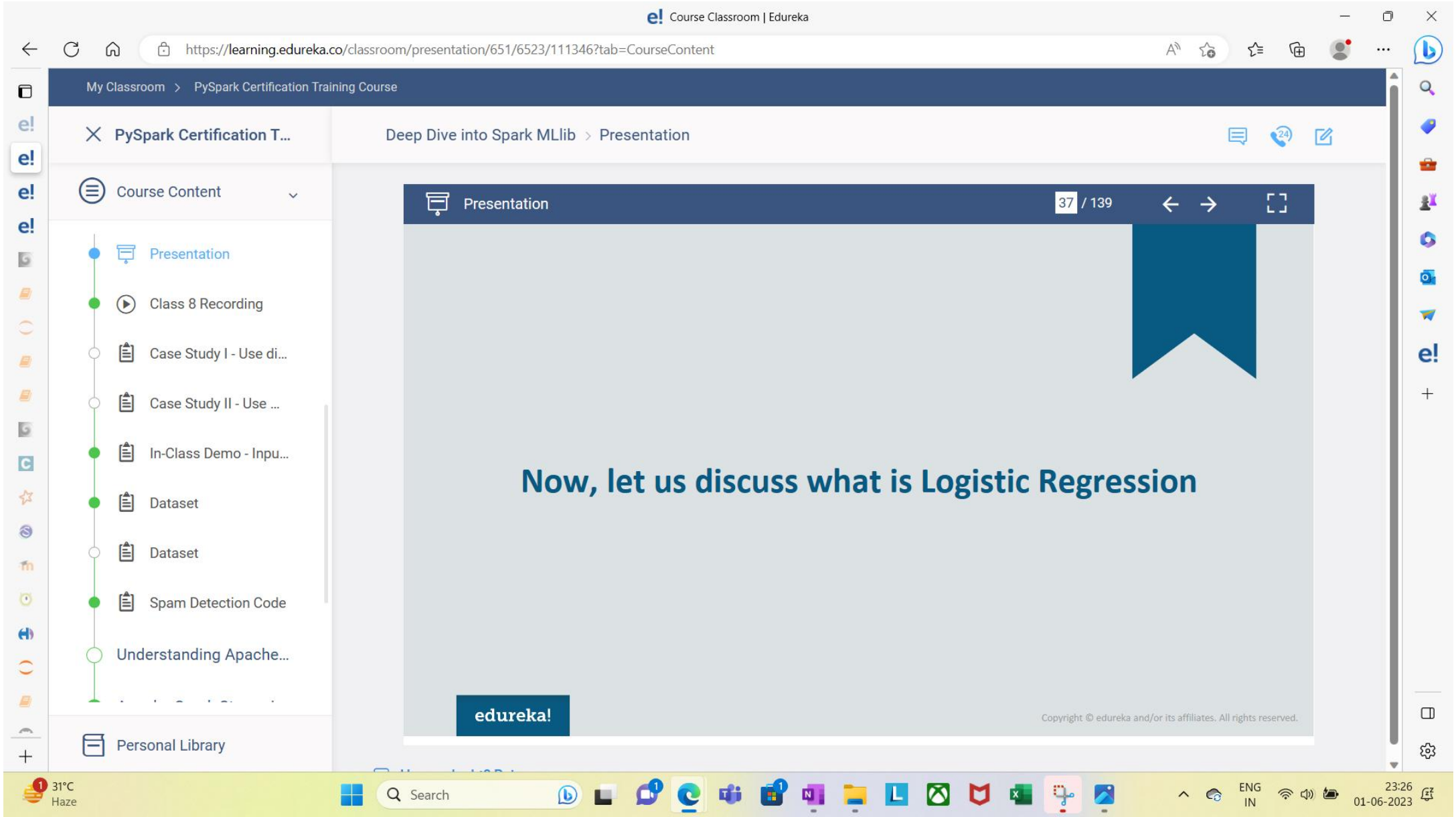
Linear line has to be clipped at 0 and 1

With linear regression, our resulting curve cannot be formulated into a single formula as we obtain three different straight lines.

We need a new way to solve this problem

Hence we came up with Logistic Regression

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
37 / 139
Now, let us discuss what is Logistic Regression
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

Presentation 38 / 139

Logistic Regression - Use Case

A patient goes for a routine check up in a hospital.
He is interested to know whether the cancer is benign or malignant

Let's make a "Logistic Regression Model" to predict whether a patient is benign or malignant
AND
Check the accuracy of the Model

edureka!

Presentation 39 / 139

What is Logistic Regression?

For Example:-

A patient's data such as sugar level, blood pressure, age, skin width, previous medical history are recorded.

edureka!

Presentation 40 / 139

What is Logistic Regression?

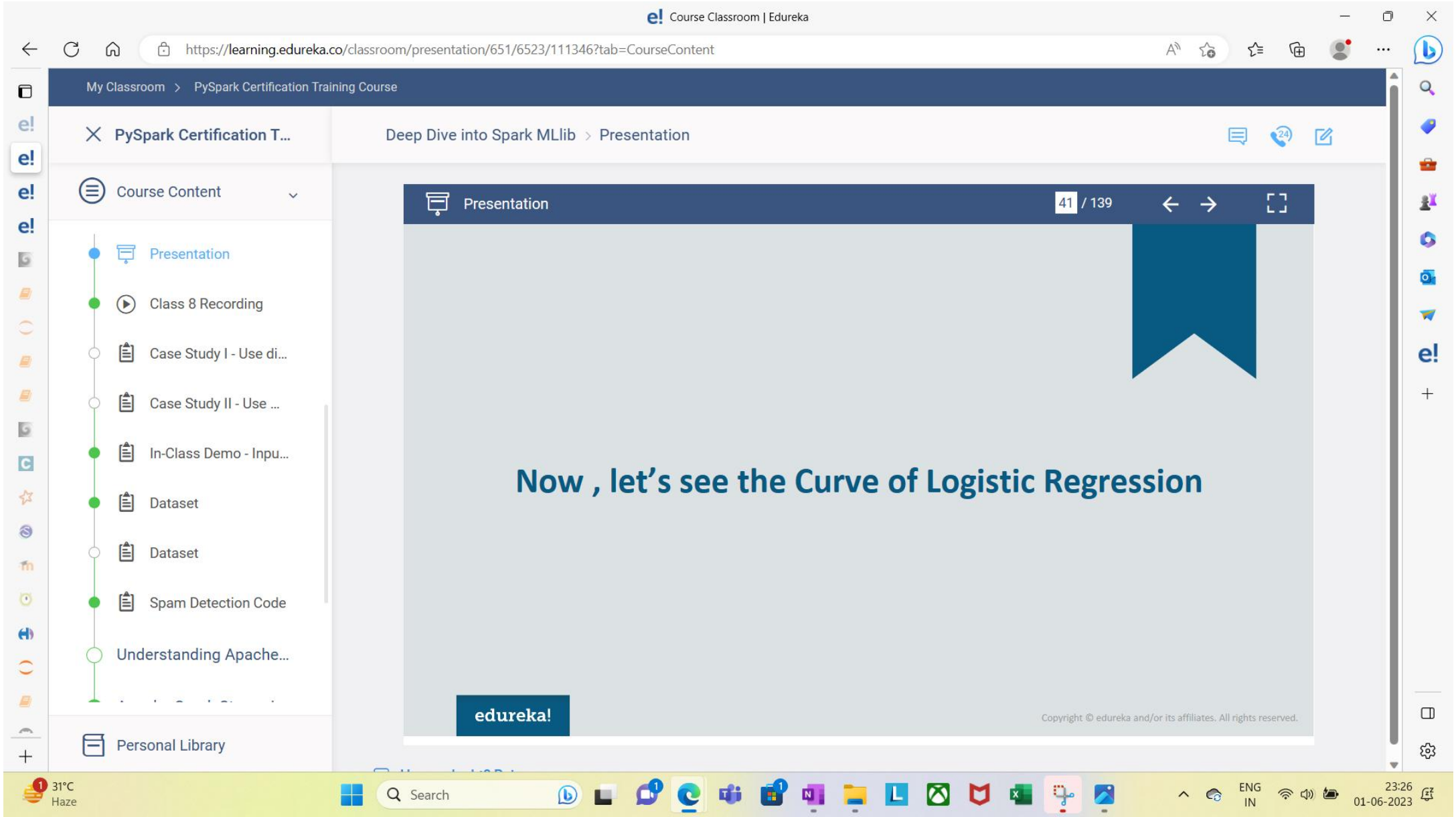
For Example:-

Doctor checks the patient data and determines the outcome of his illness and severity of illness.

The outcome(result) will be binary(0/1), i.e.
0- If malignant
1- If benign

Logistic Regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. Outcome is a binary class type.

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
41 / 139
Now , let's see the Curve of Logistic Regression
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:26
01-06-2023

Presentation 42 / 139

The Logistic Regression Curve

The Logistic Regression Curve is called as **"Sigmoid Curve"**, also known as **S-Curve**

The sigmoid function converts any value from –infinity to infinity to the discreet values 0 or 1.

How to decide whether the value is 0 or 1 from this curve?

edureka!

Presentation 43 / 139

The Logistic Regression Curve

Based on the threshold value set, we decide the output from the function

Let's take an example with threshold value 0.4

Any value above 0.4 will be rounded off to 1 and below 0.4 will be reduced to 0

Let's see how an equation is formed to imitate this function

edureka!

Presentation 44 / 139

Logistic Regression Equation

The Logistic Regression Equation is derived from the Straight Line Equation

Equation of a straight line

$Y = C + BIX1 + B2X2 + ...$ → Range is from –(infinity) to (infinity)

Let's try to reduce the Logistic Regression Equation from Straight Line Equation

$Y = C + BIX1 + B2X2 + ...$ → In Logistic equation Y can be only from 0 to 1

edureka!

Presentation 45 / 139

Logistic Regression Equation

The Logistic Regression Equation is derived from the Straight Line Equation

Equation of a straight line

$Y = C + BIX1 + B2X2 + \ldots.$ → Range is from –(infinity) to (infinity)

Let's try to reduce the Logistic Regression Equation from Straight Line Equation

$Y = C + BIX1 + B2X2 + \ldots.$ → In Logistic equation Y can be only from 0 to 1

Now , to get the range of Y between 0 and infinity, let's transform Y

$\frac{Y}{1-Y}$ } Y= 0 then 0, Y= 1 then infinity → Now, the range is between 0 to infinity

edureka!

Logistic Regression Equation

The Logistic Regression Equation is derived from the Straight Line Equation

Equation of a straight line

$Y = C + BIX1 + B2X2 + \ldots.$ → Range is from –(infinity) to (infinity)

Let's try to reduce the Logistic Regression Equation from Straight Line Equation

$Y = C + BIX1 + B2X2 + \ldots.$ → In Logistic equation Y can be only from 0 to 1

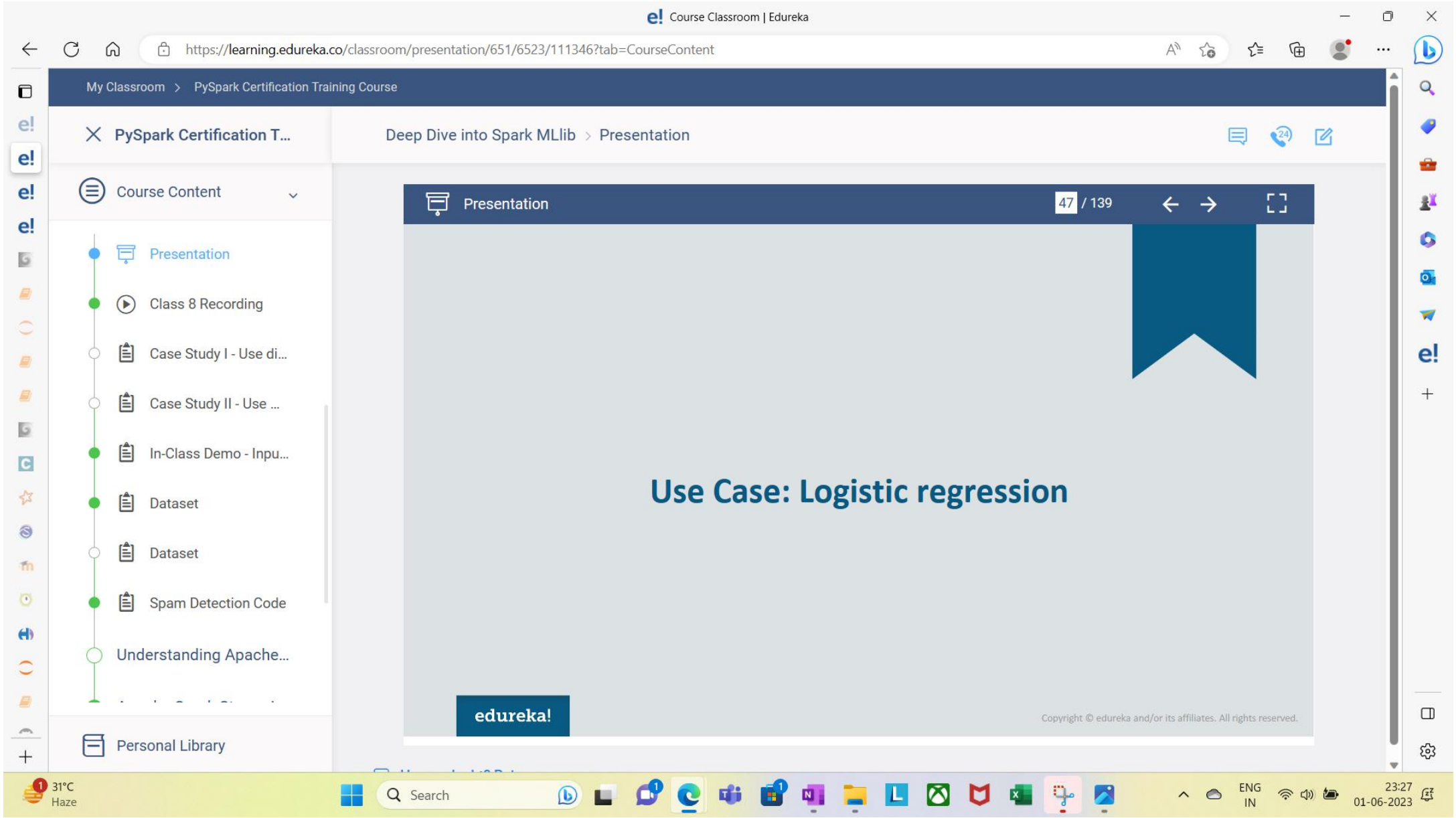
Now , to get the range of Y between 0 and infinity, let's transform Y

$\frac{Y}{1-Y}$ } Y= 0 then 0, Y= 1 then infinity → Now, the range is between 0 to infinity

Let us transform it further, to get range between –(infinity) and (infinity)

$\log\left[\frac{Y}{1-Y}\right] \Rightarrow Y = C + BIX1 + B2X2 + \ldots.$ → **Final Logistic Regression Equation**

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
47 / 139
Use Case: Logistic regression
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 48 / 139

Passenger Survival Prediction

We all know about the "Titanic" disaster. The survey shows that only 38% of the passengers survived. The reason for this massive loss of life is that the Titanic was only carrying 20 lifeboats, which was not nearly enough for the 1,317 passengers and 885 crew members aboard. It seems unlikely that all of the passengers would have had equal chances at survival. We want to write a program to predict whether a given passenger would survive the disaster

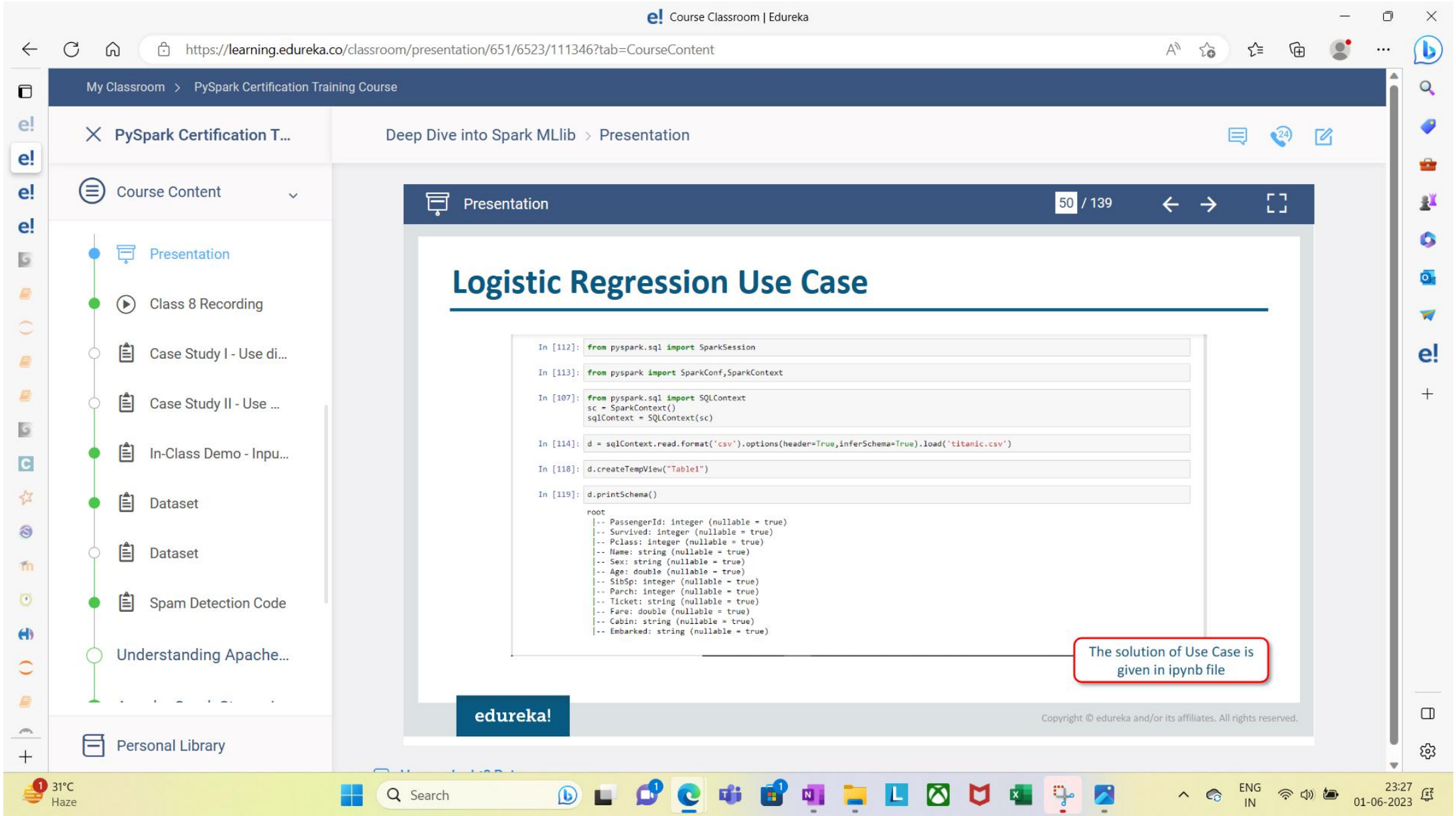
edureka!

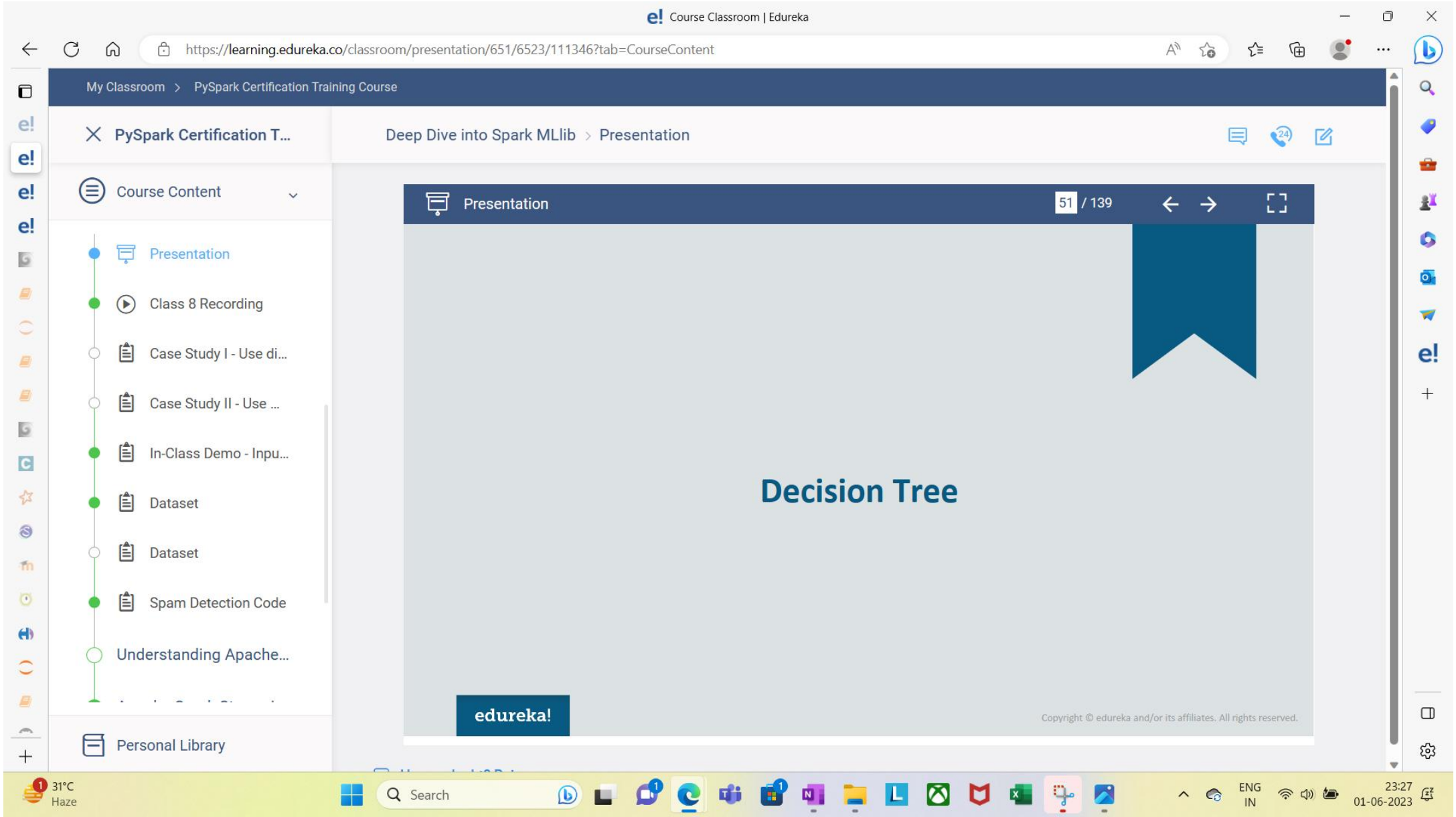
Presentation 49 / 139

Dataset Overview – Titanic.CSV

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S	
3	2	1	1	Cumings,	female	38	1	0	PC 17599	71.2833	C85	C	
4	3	1	3	Heikkinen	female	26	0	0	STON/O2.	7.925		S	
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S	
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S	
7	6	0	3	Moran, Mi	male		0	0	330877	8.4583		Q	
8	7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S	
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S	
10	9	1	3	Johnson, I	female	27	0	2	347742	11.1333		S	
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C	
12	11	1	3	Sandstron	female	4	1	1	PP 9549	16.7	G6	S	
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S	
14	13	0	3	Saundercc	male	20	0	0	A/5. 2151	8.05		S	
15	14	0	3	Andersso	male	39	1	5	347082	31.275		S	
16	15	0	3	Vestrom,	female	14	0	0	350406	7.8542		S	
17	16	1	2	Hewlett, M	female	55	0	0	248706	16		S	
18	17	0	3	Rice, Mast	male	2	4	1	382652	29.125		Q	
19	18	1	2	Williams,	male		0	0	244373	13		S	
20	19	0	3	Vander Pl	female	31	1	0	345763	18		S	
21	20	1	3	Masselma	female		0	0	2649	7.225		C	
22	21	0	2	Fynney, M	male	35	0	0	239865	26		S	
23	22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S	
24	23	1	3	McGowan	female	15	0	0	330923	8.0292		Q	

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
50 / 139
Logistic Regression Use Case
In [112]: from pyspark.sql import SparkSession
In [113]: from pyspark import SparkConf,SparkContext
In [107]: from pyspark.sql import SQLContext
sc = SparkContext()
sqlContext = SQLContext(sc)
In [114]: d = sqlContext.read.format('csv').options(header=True,inferSchema=True).load('titanic.csv')
In [118]: d.createTempView("Table1")
In [119]: d.printSchema()
root
|-- PassengerId: integer (nullable = true)
|-- Survived: integer (nullable = true)
|-- Pclass: integer (nullable = true)
|-- Name: string (nullable = true)
|-- Sex: string (nullable = true)
|-- Age: double (nullable = true)
|-- SibSp: integer (nullable = true)
|-- Parch: integer (nullable = true)
|-- Ticket: string (nullable = true)
|-- Fare: double (nullable = true)
|-- Cabin: string (nullable = true)
|-- Embarked: string (nullable = true)
The solution of Use Case is given in ipynb file
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C Haze
Search
ENG IN
23:27 01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
51 / 139
Decision Tree
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 52 / 139

Decision Tree

Let's take an example,

Given a dataset which contains weather information of last 14 days and also a variable called 'Play' which denotes whether the game was played as per weather conditions.

Now using the Decision Tree we need to predict whether the game will happen if the conditions are as given below,

Outlook	=	Rain
Humidity	=	High
Wind	=	Weak
Play	=	?

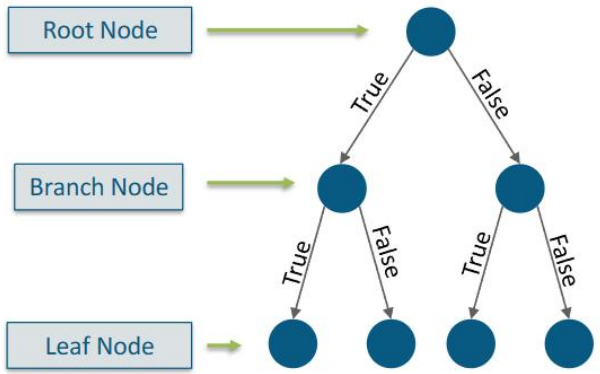
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

edureka!

Presentation 53 / 139

Decision Tree

- A decision tree is a tree-like structure in which internal node represents test on an attribute.
- Each branch represents outcome of test and each leaf node represents class label (decision taken after computing all attributes).
- A path from root to leaf represents classification rules.

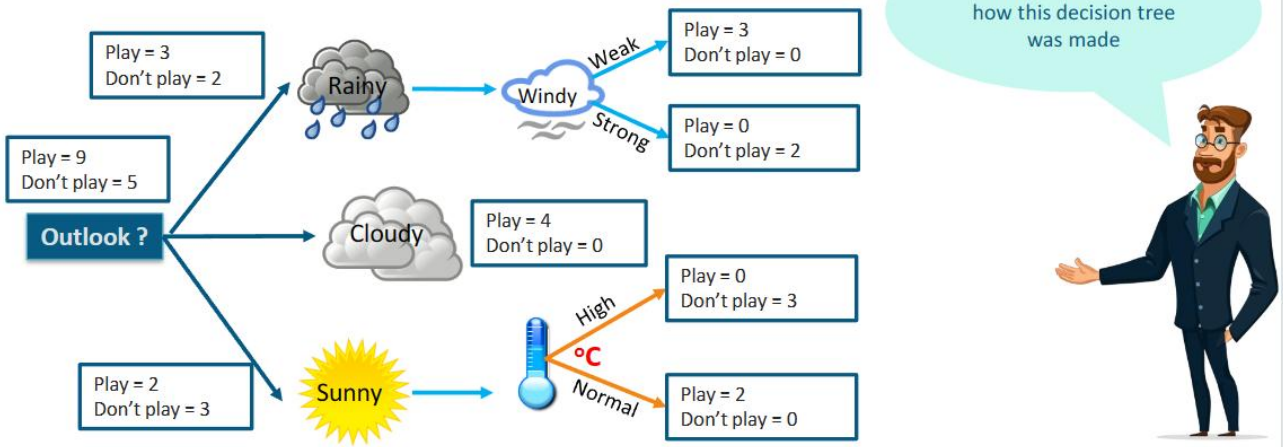


edureka!

Presentation 54 / 139

Decision Tree

A Decision Tree is made from our data by analyzing the variables. From the Decision Tree we can easily find out whether there will be game tomorrow if the conditions are rainy and less windy.



edureka!

Presentation 55 / 139

Building a Decision Tree

From each of the outlooks' we can divide the data as,

9 Yes / 5 No

Outlook

Sunny | Overcast | Rain

Day	Outlook	Humidity	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

2 Yes / 3 No

Split further

Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Pure subset

Yes(will play)

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

3 Yes / 2 No

Split further

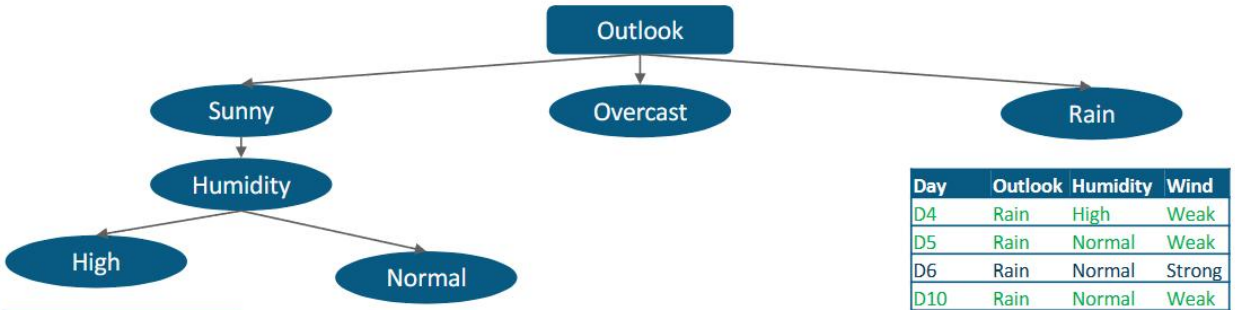
edureka!

PySpark Certification T...
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library

Presentation 56 / 139

Building a Decision Tree

We will use Humidity column to split the subset sunny further,



Day	Humidity	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Pure subset
NO(will not play)

Day	Humidity	Wind
D9	Normal	Weak
D11	Normal	Strong

Pure subset
Yes(will play)

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

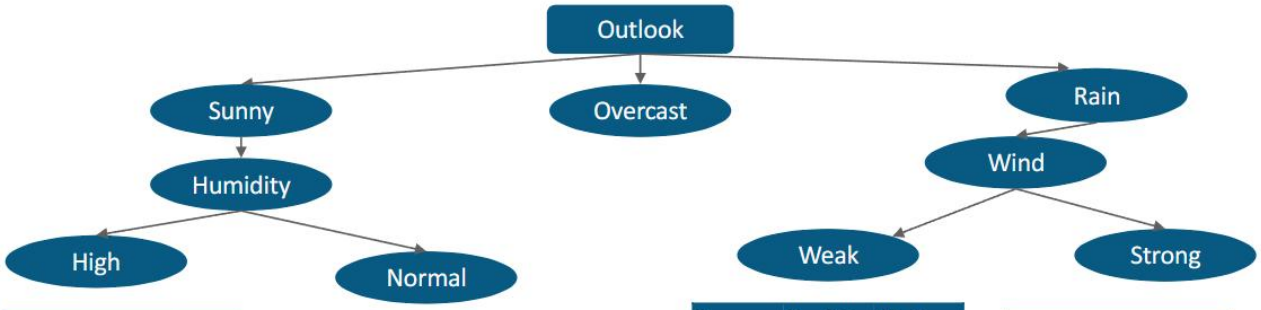
3 Yes / 2 No

Split further

edureka!

Building a Decision Tree

Similarly we will divide rain subset using Humidity,



Day	Humidity	Wind
D1	High	Weak
D2	High	Strong
D8	High	Weak

Pure subset
NO(will not play)

Day	Humidity	Wind
D9	Normal	Weak
D11	Normal	Strong

Pure subset
Yes(will play)

Day	Humidity	Wind
D4	High	Weak
D5	Normal	Weak
D10	Normal	Weak

Pure subset
Yes(will play)

Day	Humidity	Wind
D6	Normal	Strong
D14	High	Strong

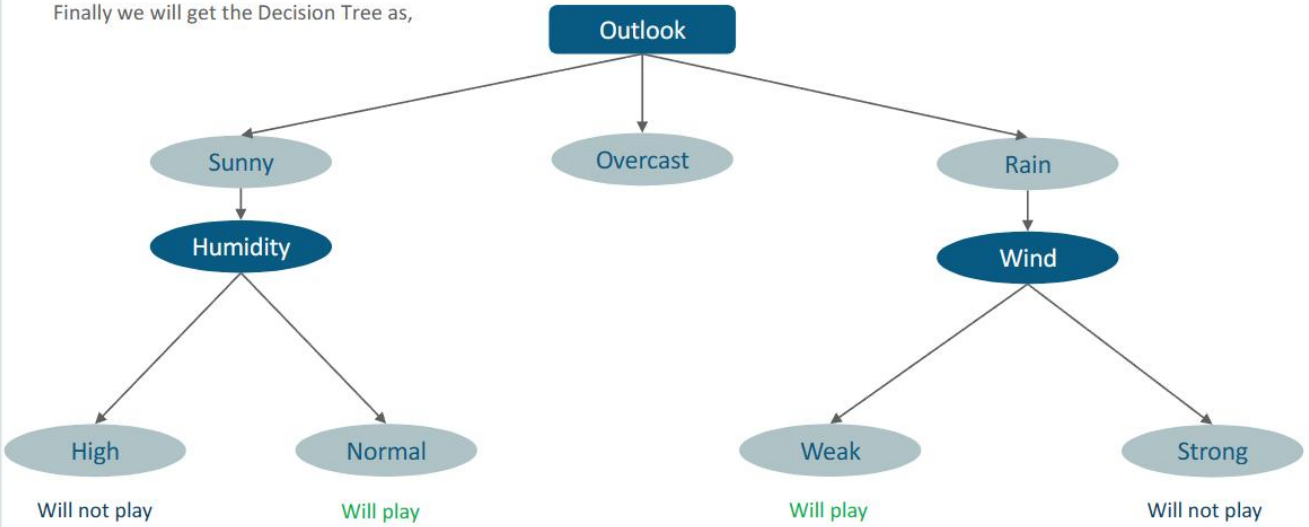
Pure subset
NO(will not play)

edureka!

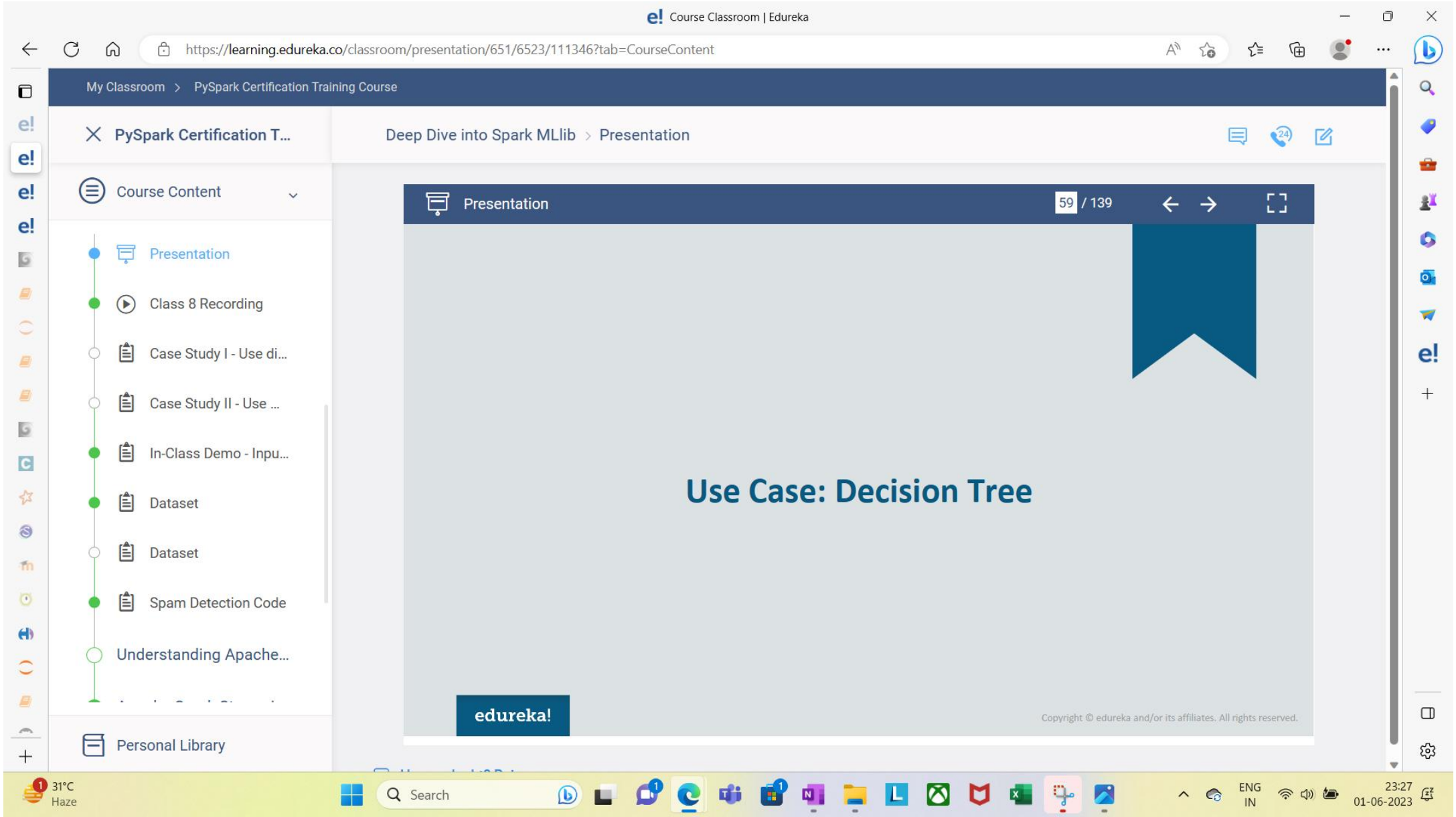
Presentation 58 / 139

Building a Decision Tree

Finally we will get the Decision Tree as,



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
59 / 139
Use Case: Decision Tree
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 60 / 139

Use Case: Decision Tree

- There is one shop named "**Riders**", who gives bikes on rent. The bike rental count differs according to season, year, month, hour, weather etc.
- We want to predict bike rental counts (per hour) from information such as day of the week, weather, season, etc. Having good predictions of customer demand allows a business or service to prepare and increase supply as needed

edureka!

Presentation 61 / 139

Use Case: Decision Tree

The parameters which decides the bike rental count are given below:

Date, Month, Year

Season, Weather

Temperature, Humidity

Holiday/ Weekday

edureka!

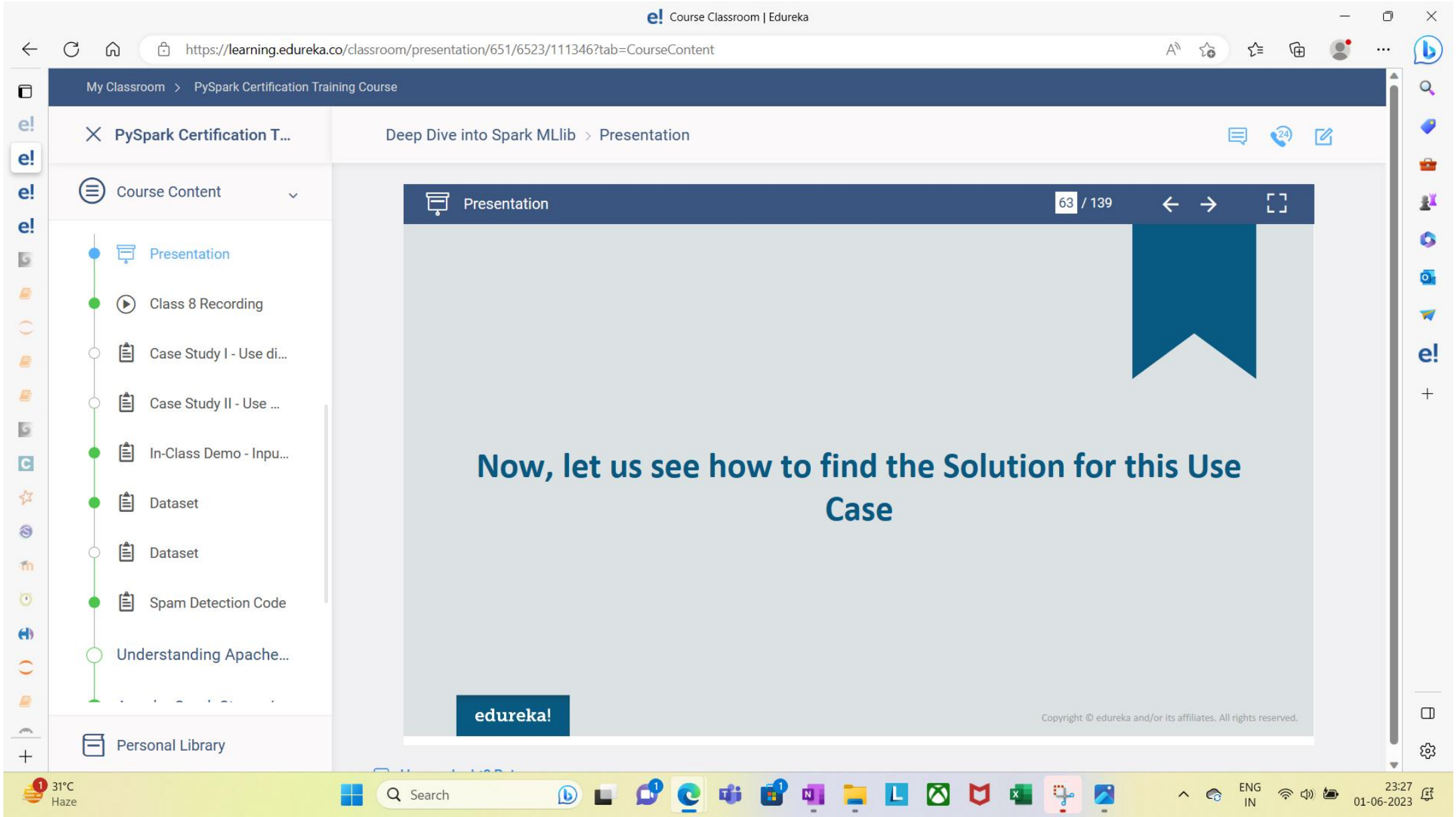
Presentation 62 / 139

Use Case: Decision Tree

The parameters which decides the bike rental count are given below:

- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
63 / 139
Now, let us see how to find the Solution for this Use Case
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 64 / 139

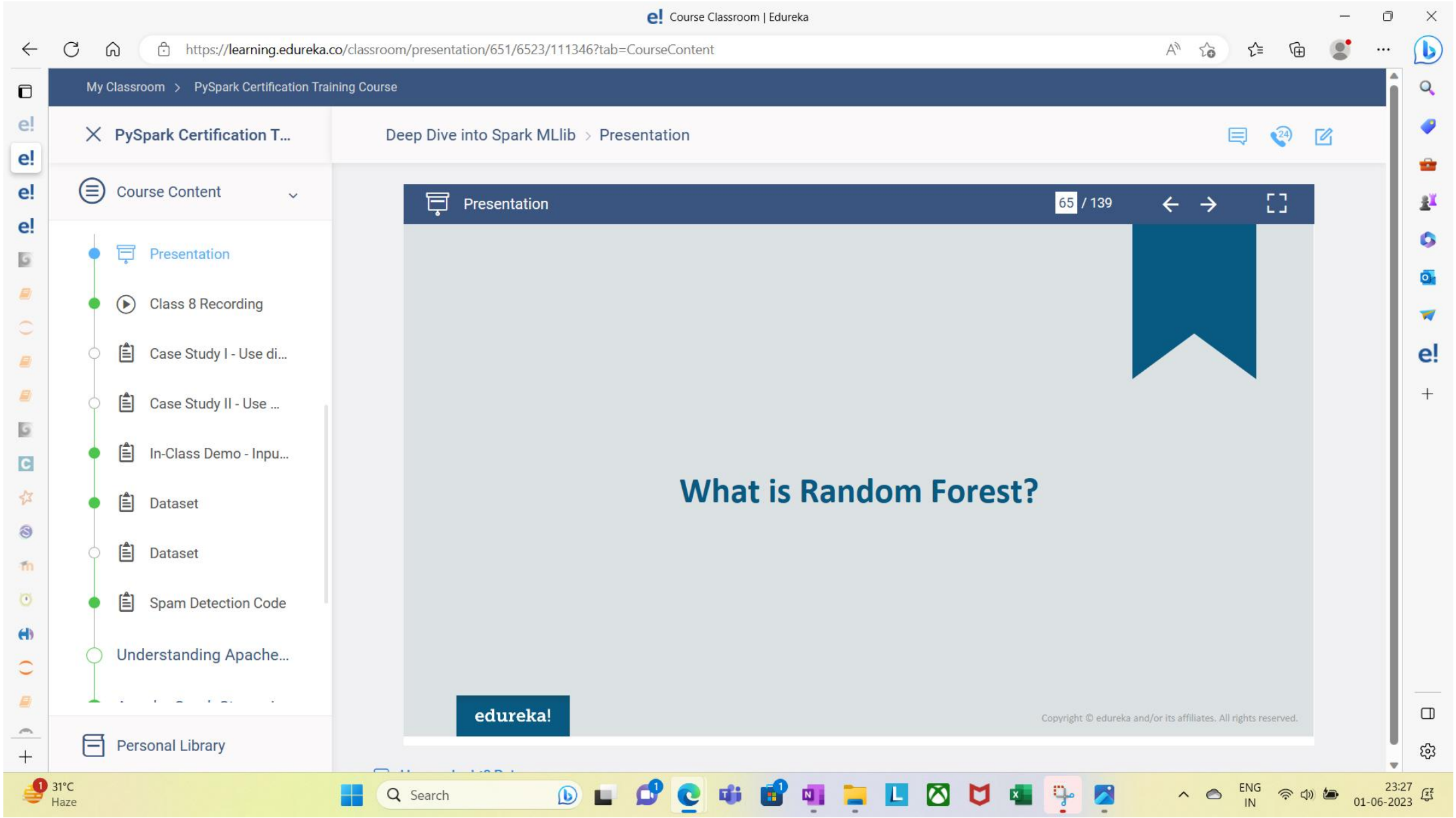
Use Case Solution: Decision Tree

```
In [1]: from pyspark import SparkConf, SparkContext
        from pyspark.sql import SQLContext

In [2]: sc= SparkContext()
        sqlContext = SQLContext(sc)

In [3]: df = sqlContext.read.format('csv').option("header", 'true').load("hdfs://nameservice1/user/edureka_294428/bikeSharing.csv")

In [4]: df.cache()

Out[4]: DataFrame[instant: string, dteday: string, season: string, yr: string, mnth: string, hr: string, holiday: string, weekday: string, workingday: string, weathersit: string, temp: string, atemp: string, hum: string, windspeed: string, casual: string, registered: string, cnt: string]

In [5]: df.show()
```

instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
6	2011-01-01	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
7	2011-01-01	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
8	2011-01-01	1	0	1	7	0	6	0	1	0.2	0.2576	0.86				
9	2011-01-01	1	0	1	8	0	6	0	1	0.24	0.2879	0.75				

The Solution of this Use case is given in ipynb file

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
65 / 139
What is Random Forest?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

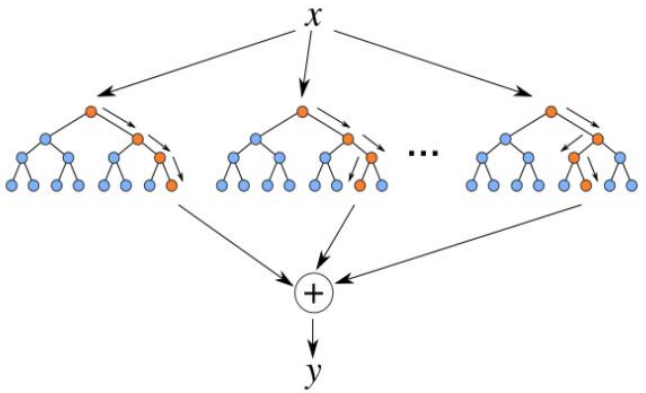
Presentation 66 / 139

What is Random Forest?

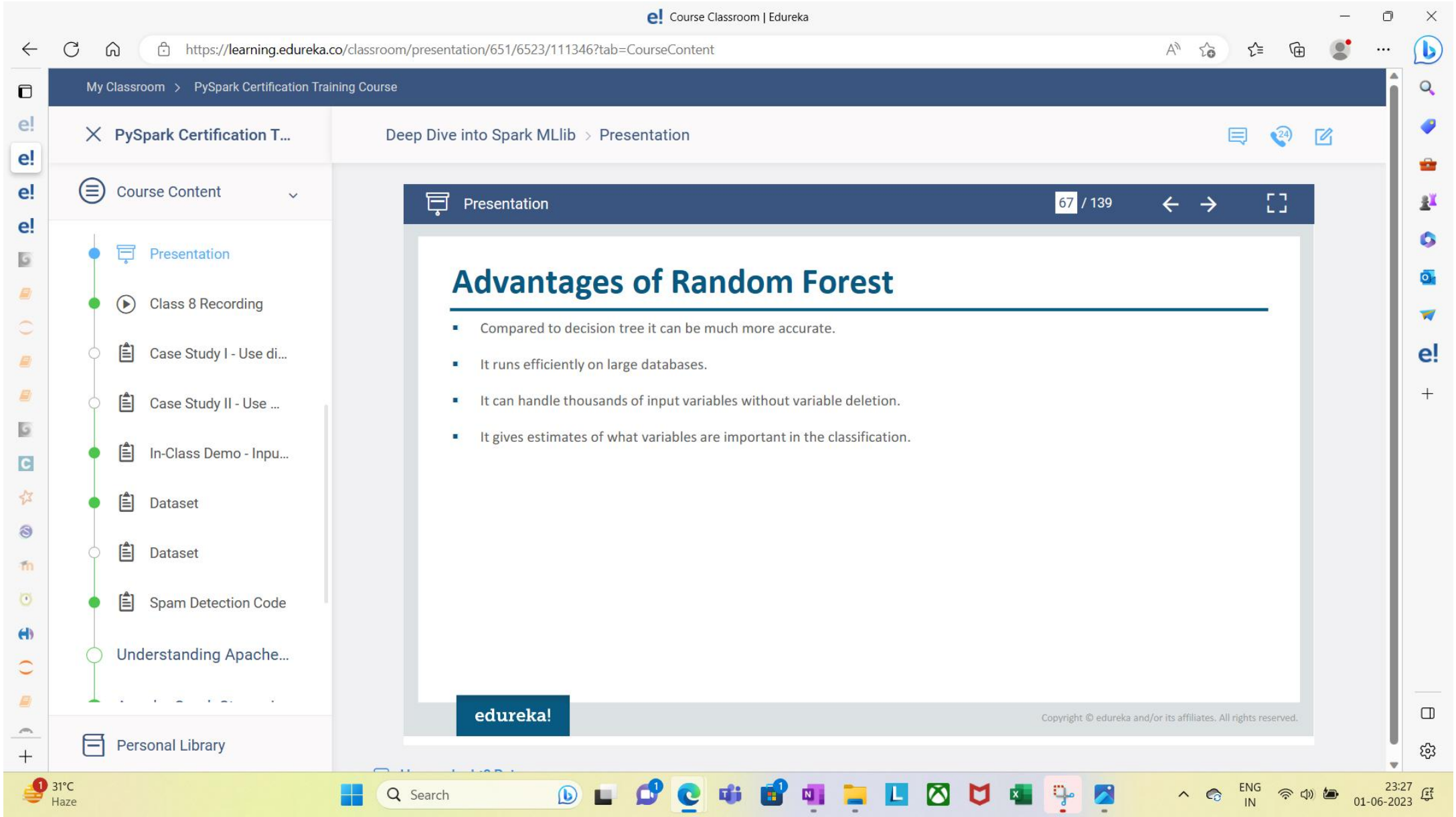
Random Forest is an ensemble classifier made using many decision tree models

What are ensemble models?

- Ensemble models combine the results from different models
- The result from an ensemble model is usually better than the result from one of the individual models
- Every tree votes for one class, the final decision is based upon majority of votes



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
67 / 139
Advantages of Random Forest
Compared to decision tree it can be much more accurate.
It runs efficiently on large databases.
It can handle thousands of input variables without variable deletion.
It gives estimates of what variables are important in the classification.
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 68 / 139

Random Forest - Example

Let's take our earlier 'weather data' and see how random forest works:

The first step in Random forest is that it will divide the data into smaller subsets.

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Note:

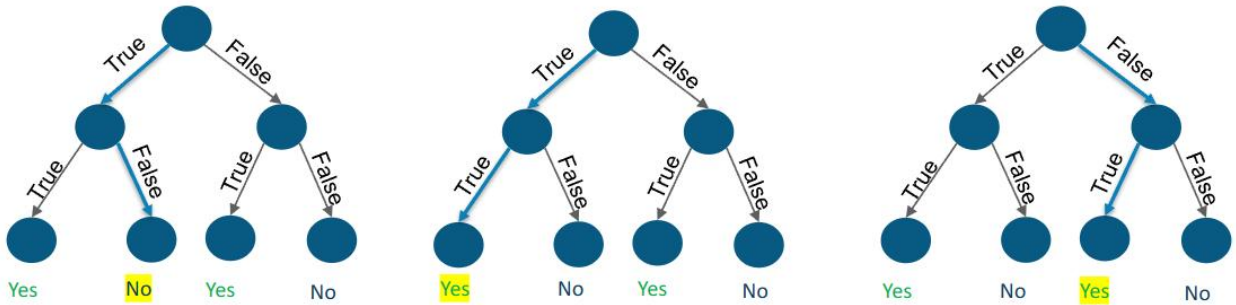
Every subsets need not be distinct, some subsets may overlap.

edureka!

Presentation 69 / 139

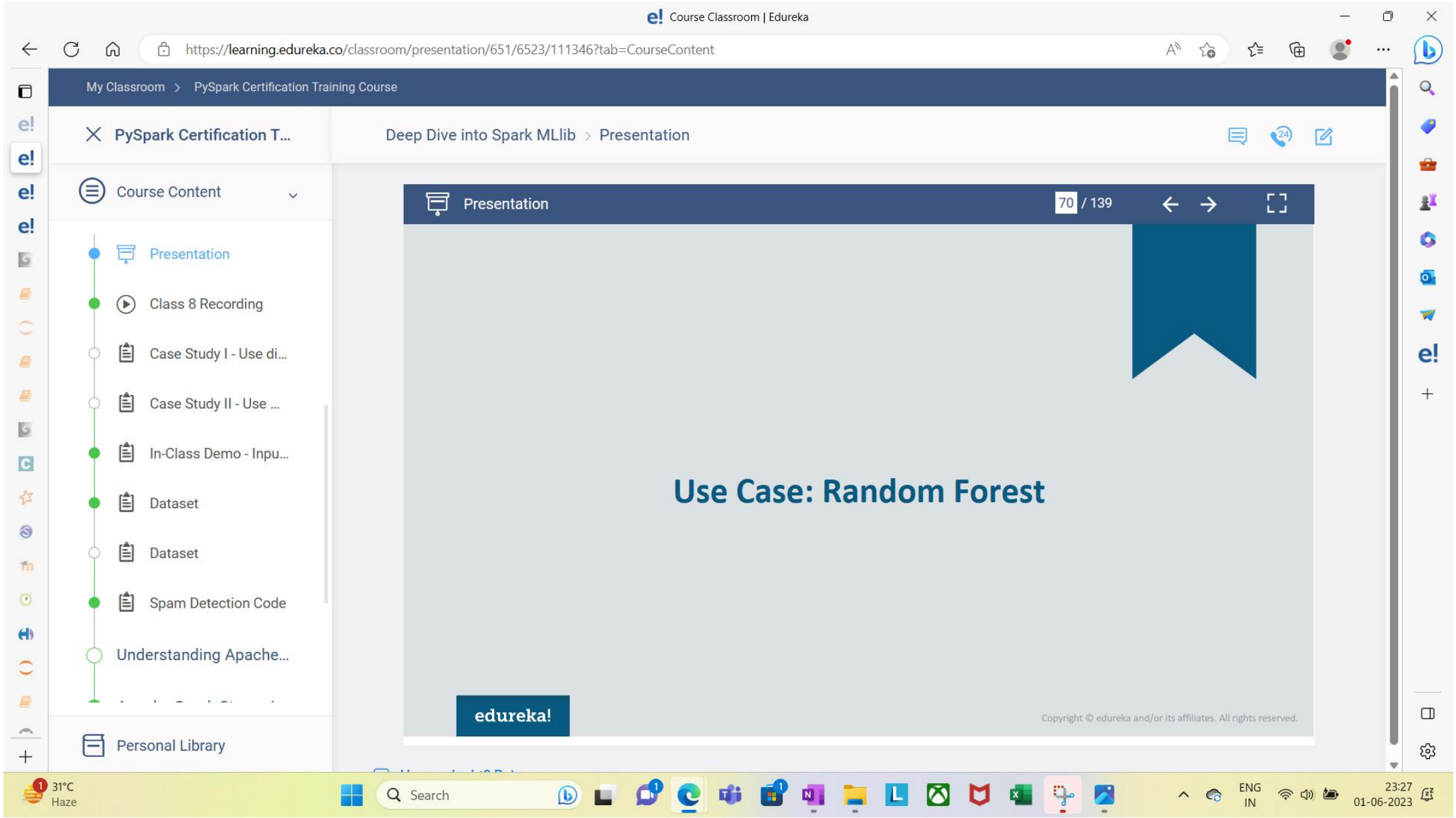
Random Forest - Example

For each subset a 'Decision tree' is made,



- Output of each tree is predicted, suppose here the first two decision trees predicted that game will happen and the third one predicted the game wont happen.
- Then based on the number of votes final output is selected, so overall predicted outcome is that the game will happen.

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
70 / 139
Use Case: Random Forest
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

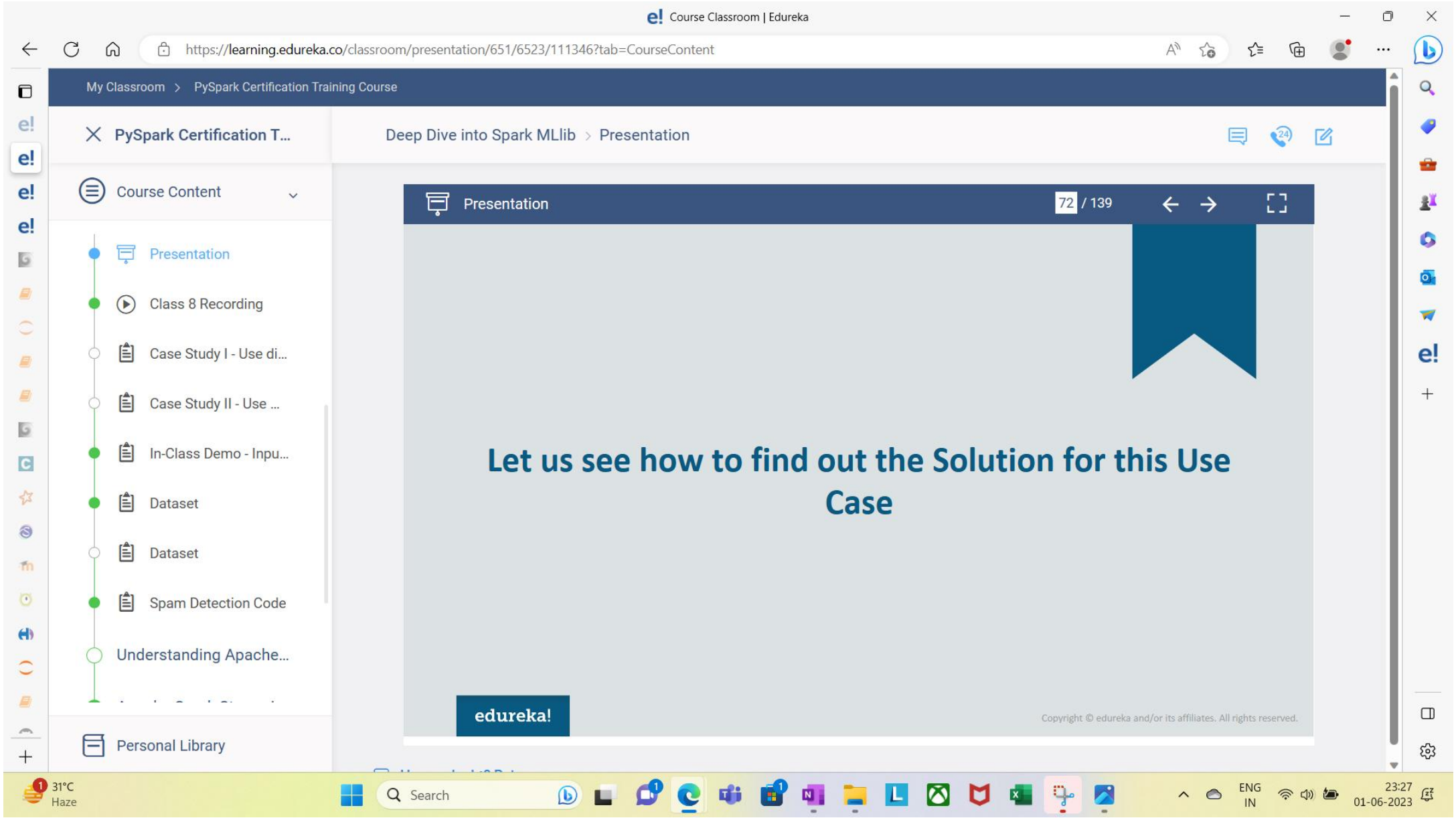
Presentation 71 / 139

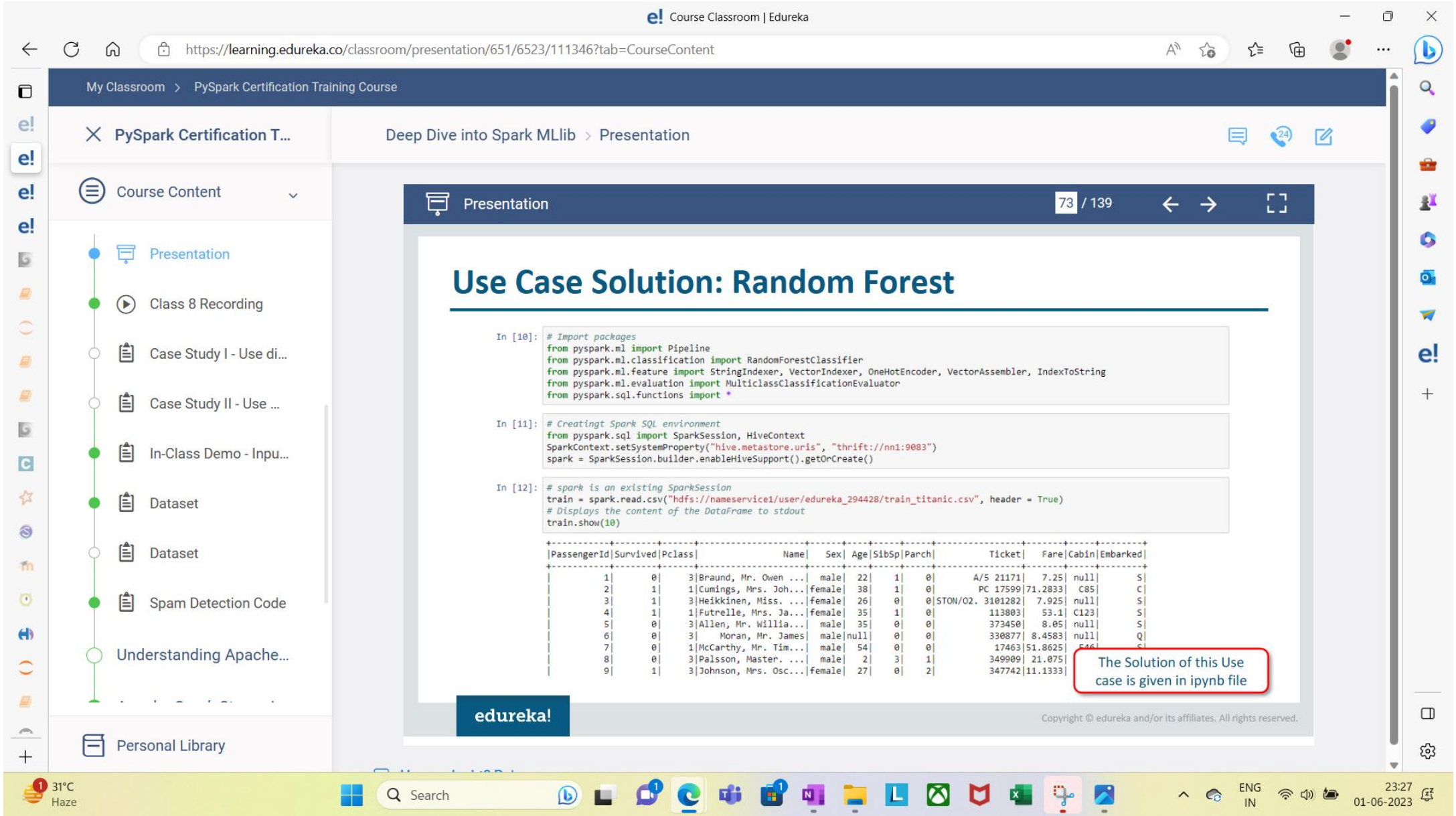
Use Case: Random Forest

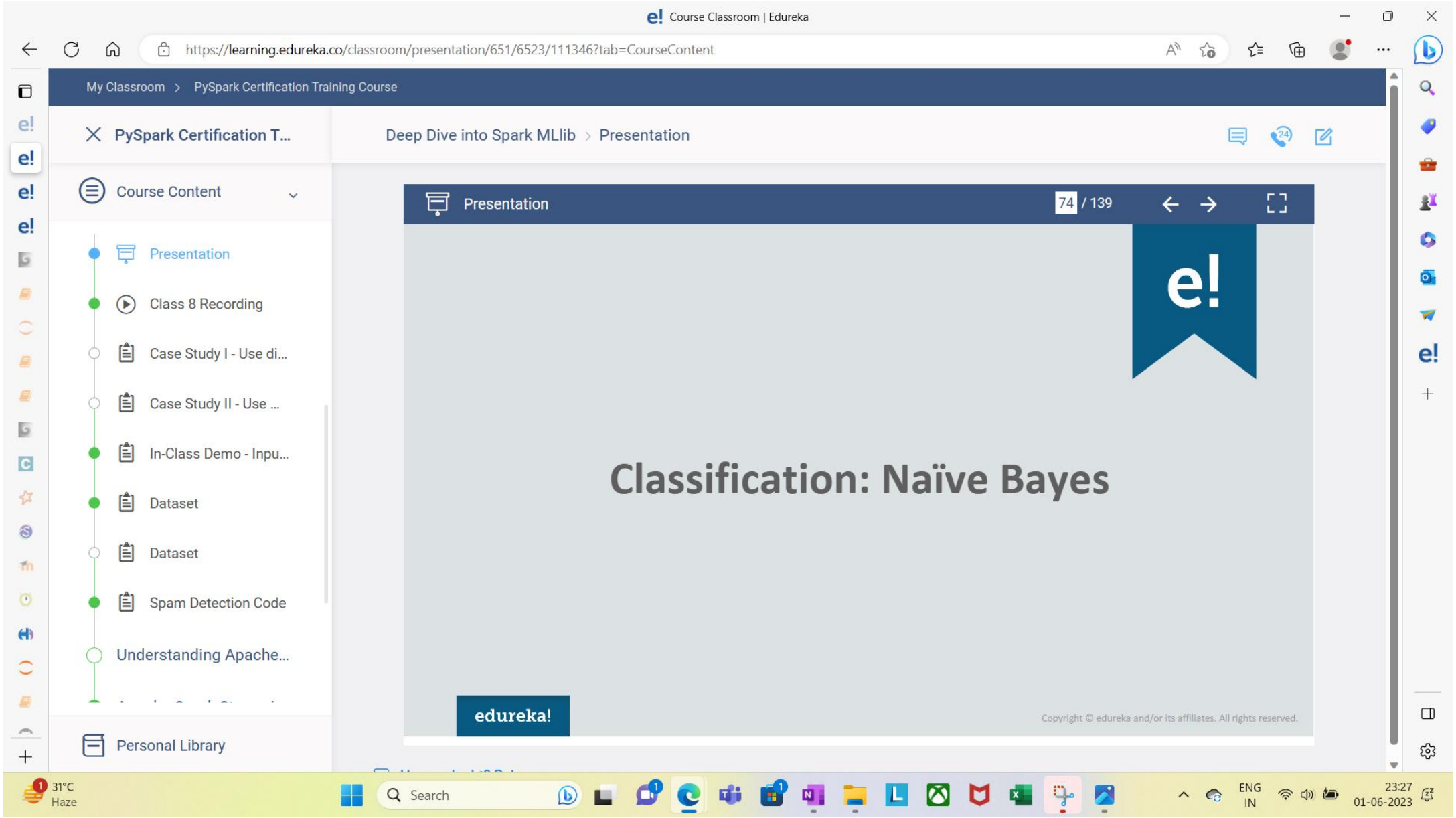
We all know about the "Titanic" disaster. The survey shows that only 38% of the passengers survived. The reason for this massive loss of life is that the Titanic was only carrying 20 lifeboats, which was not nearly enough for the 1,317 passengers and 885 crew members aboard. It seems unlikely that all of the passengers would have had equal chances at survival. We want to write a program to predict whether a given passenger would survive the disaster

The "**Random Forest**" classification algorithm will create a multitude of trees for the data set using different random subsets of the input variables, and will return whichever prediction was returned by the most trees

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
72 / 139
Let us see how to find out the Solution for this Use Case
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
73 / 139
Use Case Solution: Random Forest
In [10]:
# Import packages
from pyspark.ml import Pipeline
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.feature import StringIndexer, VectorIndexer, OneHotEncoder, VectorAssembler, IndexToString
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.sql.functions import *
In [11]:
# Creatingt Spark SQL environment
from pyspark.sql import SparkSession, HiveContext
SparkContext.setSystemProperty("hive.metastore.uris", "thrift://nn1:9083")
spark = SparkSession.builder.enableHiveSupport().getOrCreate()
In [12]:
# spark is an existing SparkSession
train = spark.read.csv("hdfs://nameservice1/user/edureka_294428/train_titanic.csv", header = True)
# Displays the content of the DataFrame to stdout
train.show(10)
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass| Name| Sex| Age|SibSp|Parch| Ticket| Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
1	0	3	Braund, Mr. Owen ...	male	22	1	0	A/5 21171	7.25	null	S
2	1	1	Cumings, Mrs. Joh...	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. ...	female	26	0	0	STON/O2. 3101282	7.925	null	S
4	1	1	Futrelle, Mrs. Ja...	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35	0	0	373450	8.05	null	S
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.4583	null	Q
7	0	1	McCarthy, Mr. Tim...	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. ...	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Osc...	female	27	0	2	347742	11.1333		S
The Solution of this Use case is given in ipynb file
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C Haze
Search
ENG IN
23:27 01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
74 / 139
Classification: Naïve Bayes
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 75 / 139

Classification: Naïve Bayes

Based on the **Bayes Theorem:** Used to solve classification problems using probability

Why is Naïve Bayes called **Naïve**?

- Considers each predictor variable to be independent of any other variable in the model
- Used for **News Categorization, Spam Filtering, Weather Prediction, Medical Diagnosis, etc**

edureka!

Have a doubt? Raise a query.

Presentation 76 / 139

Advantages and Disadvantages: Naïve Bayes

Advantages	Disadvantages
Works quickly and saves time	Error rate in classification decision
Suitable for solving multi-class prediction problems	Faces Zero Frequency Issue
Performs better than models when this assumption holds true	Assumes that predictors are independent

edureka!

Have a doubt? Raise a query.

Presentation 76 / 139

Advantages and Disadvantages: Naïve Bayes

Advantages	Disadvantages
Works quickly and saves time	Error rate in classification decision
Suitable for solving multi-class prediction problems	Faces Zero Frequency Issue
Performs better than models when this assumption holds true	Assumes that predictors are independent

edureka!

Have a doubt? Raise a query.

Presentation 77 / 139

ALICE's QUESTION

1. Naïve Bayes Algorithm assumes that all the features in a data set are equally important and independent. True or False

A> True

B> False

C> Can't Say

D> None of the above

edureka!

Have a doubt? Raise a query.

Presentation 78 / 139

ALICE's QUESTION

1. Naïve Bayes Algorithm assumes that all the features in a data set are equally important and independent. True or False

A> True

B> False

C> Can't Say

D> None of the above

Explanation: This assumption is called Conditional Independence where it assumes that all the features are equally important and independent in a data set.

edureka!

Have a doubt? Raise a query.

Presentation 79 / 139

ALICE's QUESTION

3. You are provided with a dataset of different flowers containing their petal lengths and colour. Build a model to predict the type of flower for given petal lengths and colour. What type of problem is this?

A> Clustering

B> Regression

C> Classification

D> None of the above

edureka!

Have a doubt? Raise a query.

Presentation 80 / 139

ALICE's QUESTION

3. You are provided with a dataset of different flowers containing their petal lengths and colour. Build a model to predict the type of flower for given sepal lengths and colour. What type of problem is this?

A> Clustering

B> Regression

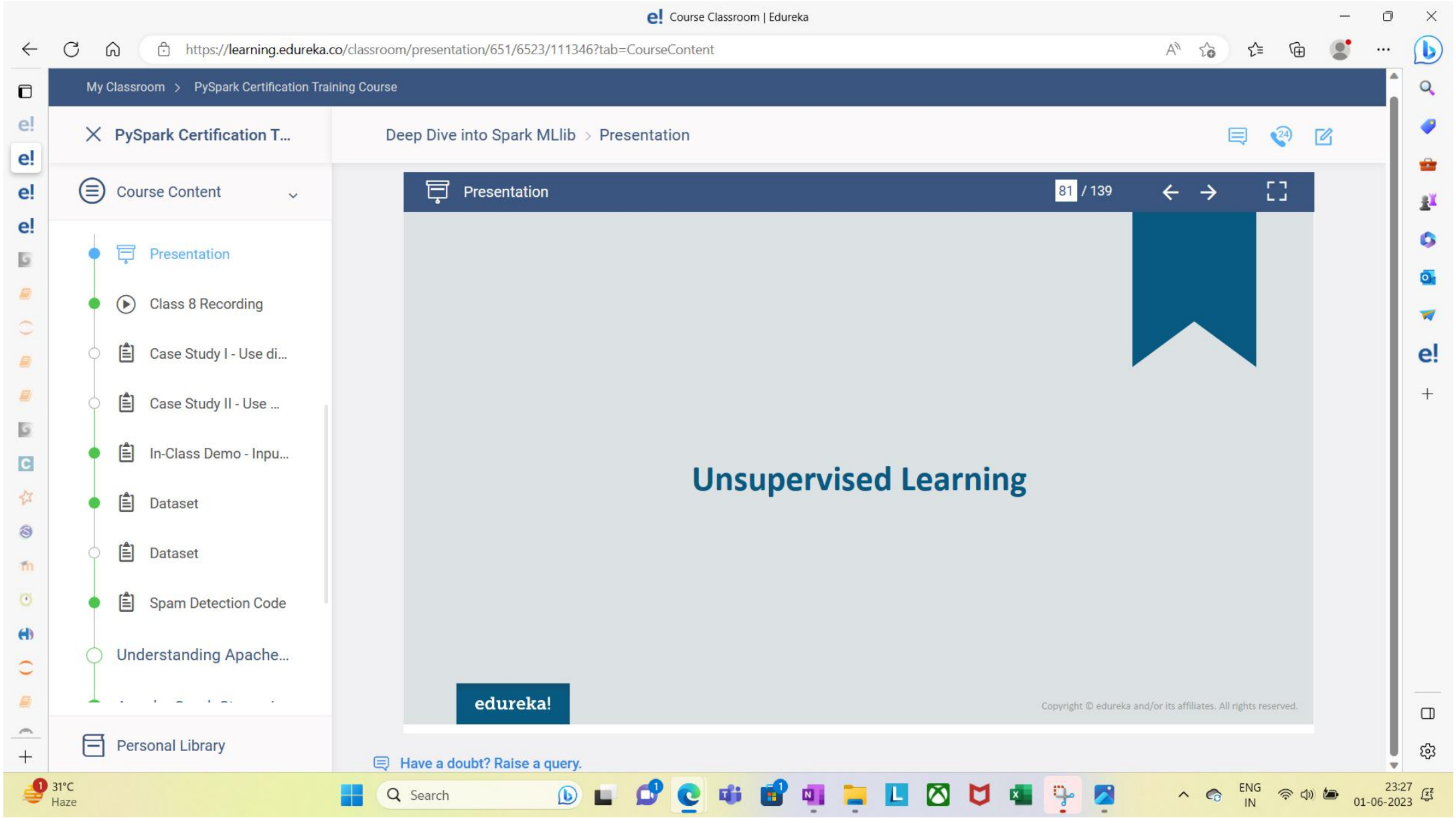
C> Classification

D> None of the above

Explanation: Here, we are classifying the type of flowers based on two features i.e. sepal lengths and colour

edureka!

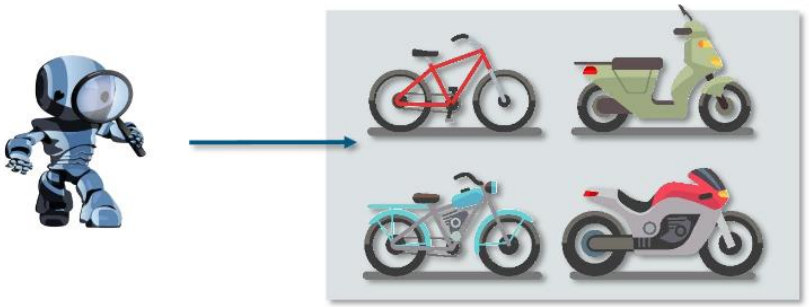
Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
81 / 139
Unsupervised Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 82 / 139

Unsupervised Learning

- Sometimes the given data is unstructured and unlabeled. So it becomes difficult to classify that data in different categories
- Unsupervised learning helps to solve this problem. This learning is used to cluster the input data in classes on the basis of their statistical properties
- Example: We can cluster different bikes based upon their speed limit, acceleration, average



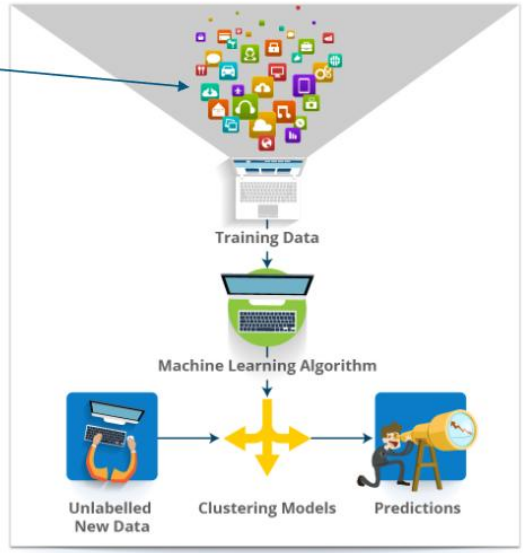
edureka!

Have a doubt? Raise a query.

Presentation 83 / 139

Unsupervised Learning – Process Flow

Training data is collection of information without any label



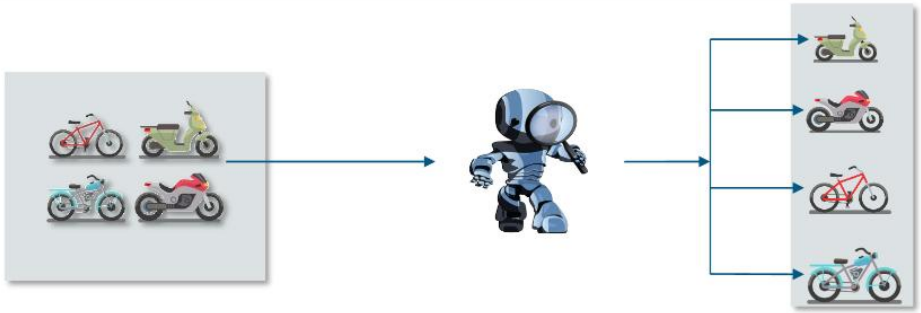
edureka!

Have a doubt? Raise a query.

Presentation 84 / 139

Unsupervised Learning - Example

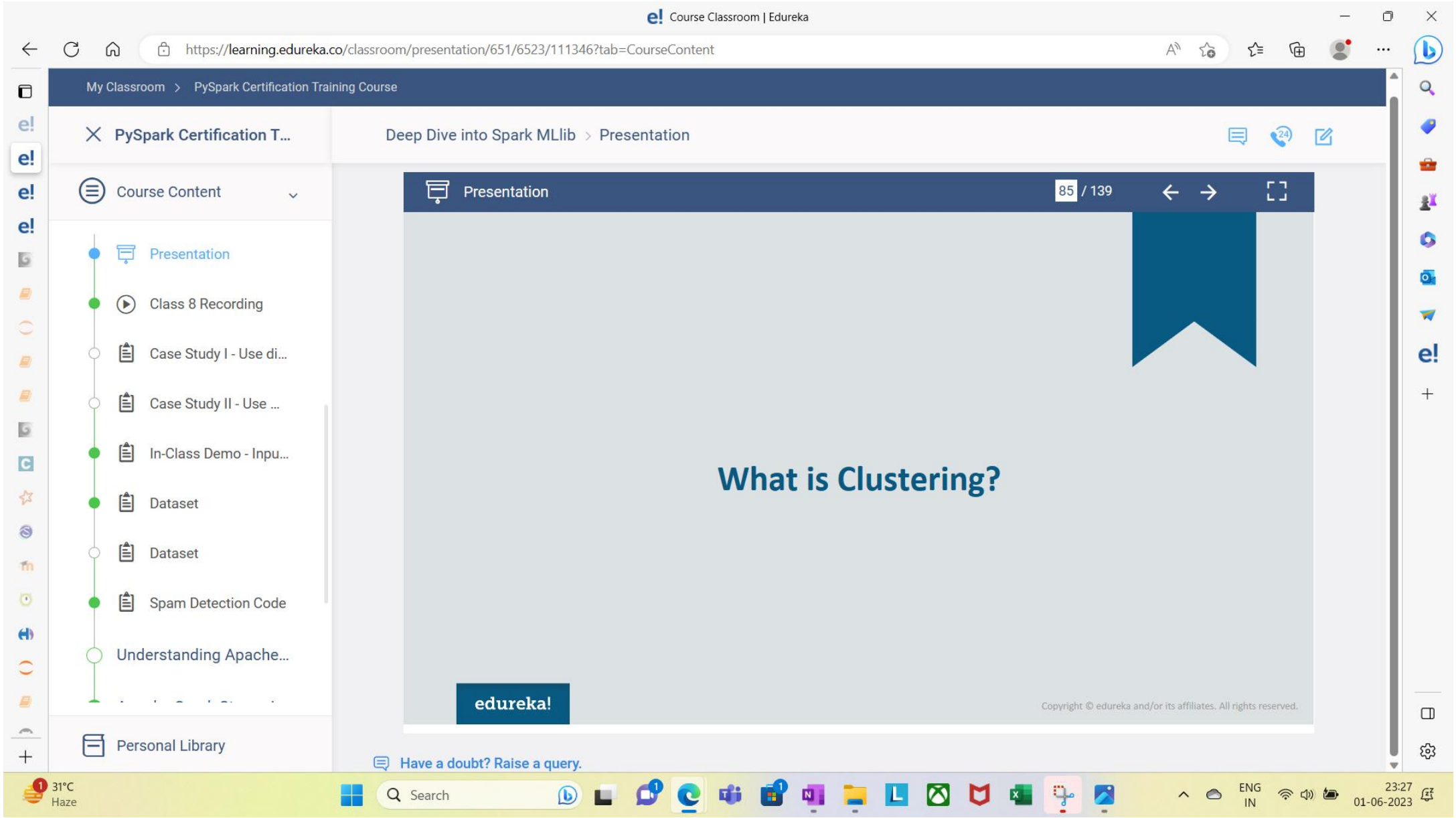
1. A set of *Bike* information is first fed into the system
2. The system identifies different Bike using features like color, size, speed limit, average etc, and it categorizes them
3. When a new *Bike* is shown, it analyses its features and puts it into the category having similar featured items



Correctness of groups depends on what attributes have been used to classify the bikes

edureka!

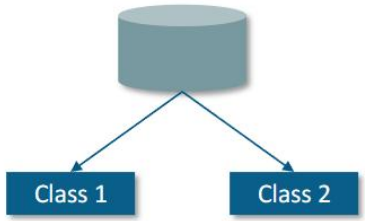
Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
85 / 139
What is Clustering?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 86 / 139

What is Clustering?

Clustering means grouping of objects based on the information found in the data, describing the objects or their relationship



The goal is that objects in one group will be similar to one other and different from objects in another group

edureka!

Have a doubt? Raise a query.

Presentation 87 / 139

Clustering

Class Student

Class Teacher

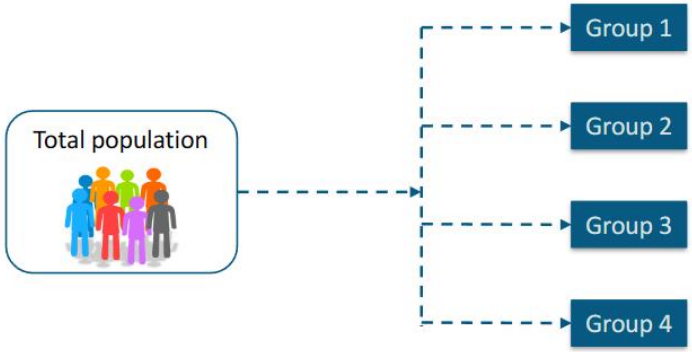
edureka!

Have a doubt? Raise a query.

Presentation 88 / 139

Clustering

- The objects in group 1 should be as similar as possible
- But there should be much difference between an object in group 1 and group 2
- The attributes of the objects are allowed to determine which objects should be grouped together



edureka!

Have a doubt? Raise a query.

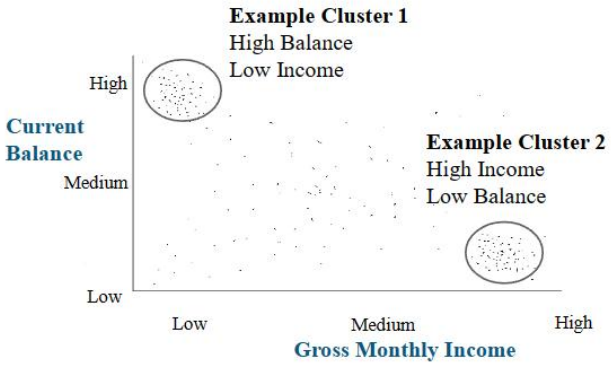
Presentation 89 / 139

Clustering

Basic concepts of Cluster Analysis using two variables

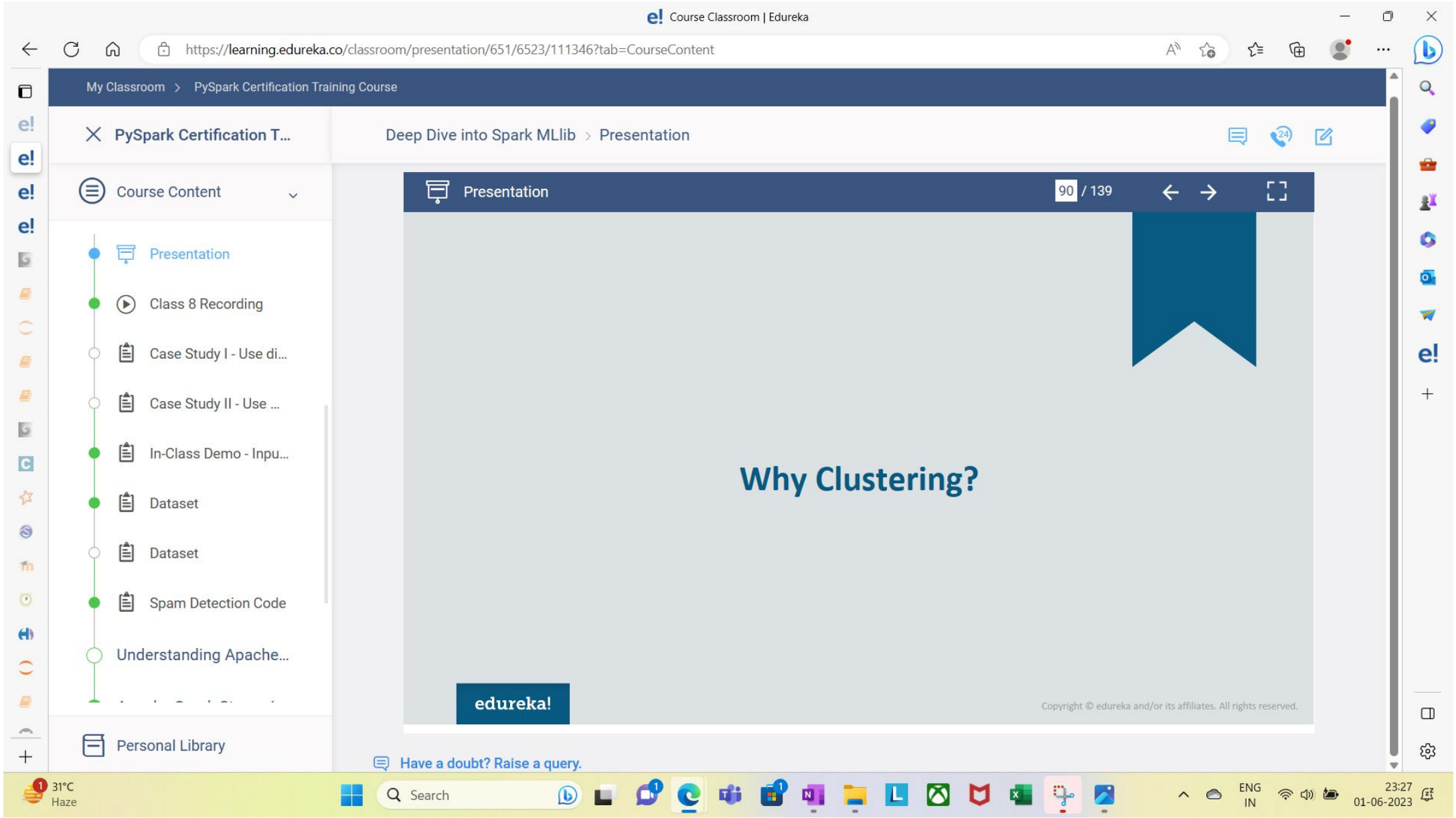
Cluster 1 and Cluster 2 are being differentiated by Income and Current Balance

- The objects in Cluster 1 have similar characteristics (High Income and Low balance)
- Also the objects in Cluster 2 have the same characteristic (High Balance and Low Income)
- But there are much differences between an object in Cluster 1 and an object in Cluster 2



edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
90 / 139
Why Clustering?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

Presentation 91 / 139

Why Clustering?

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data

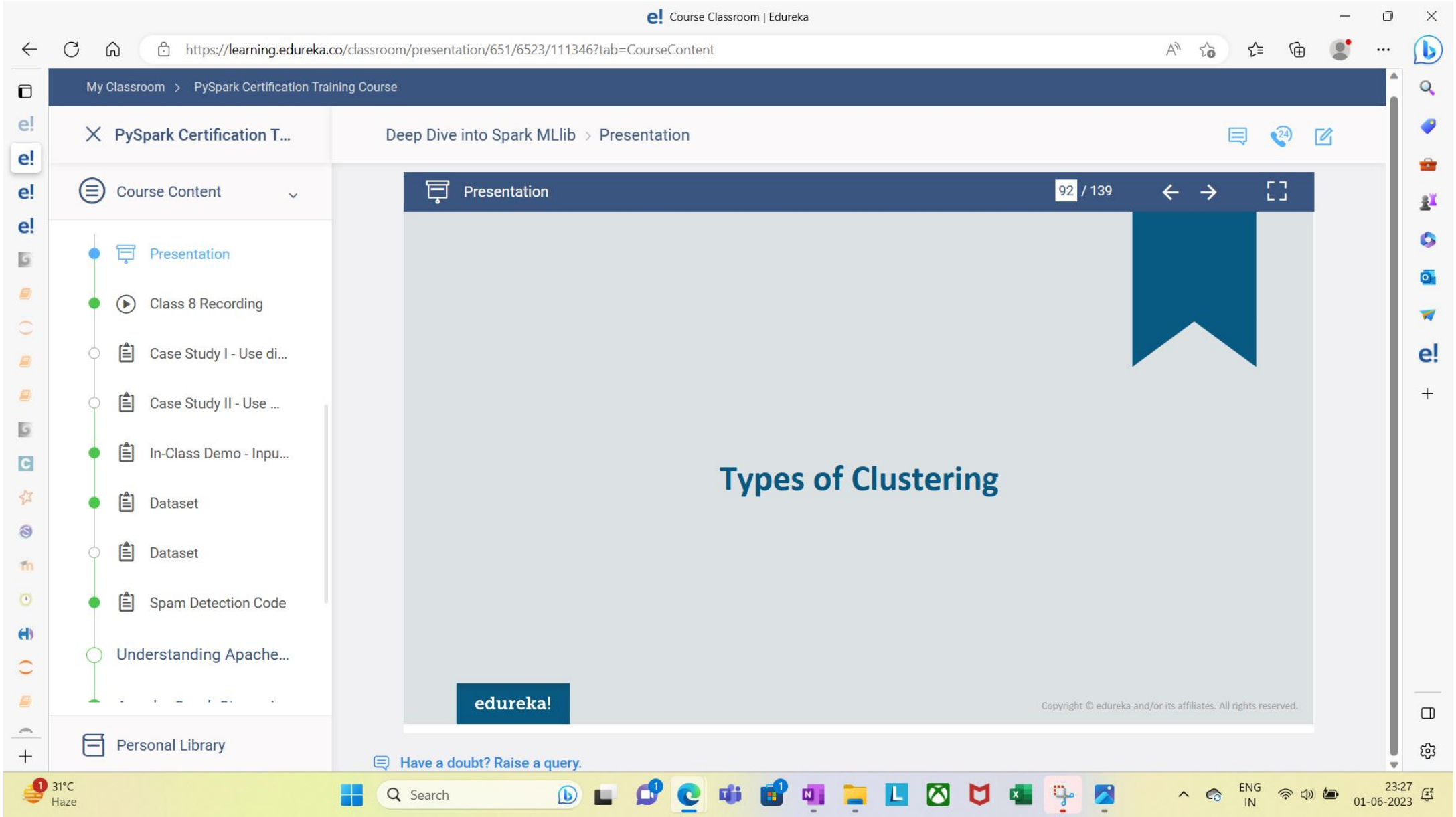
Organizing data into clusters shows internal structure of the data

Sometimes partitioning is the goal

The purpose of clustering algorithm is to make sense of, and extract value from large sets of structured and unstructured data

edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
92 / 139
Types of Clustering
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:27
01-06-2023

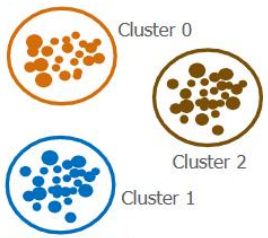
Presentation 93 / 139

Classification of Clustering

Clustering

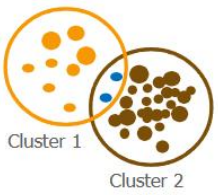
Exclusive Clustering

Here, an item belongs exclusively to one cluster, not several. K-means does this sort of exclusive clustering.



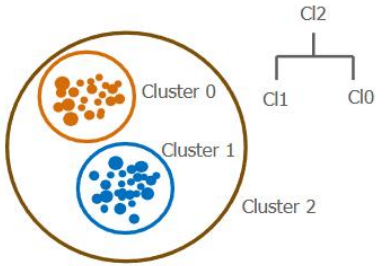
Overlapping Clustering

Here, an item can belong to multiple clusters and its degree of association with each cluster is shown. fuzzy/c-means is of this type.



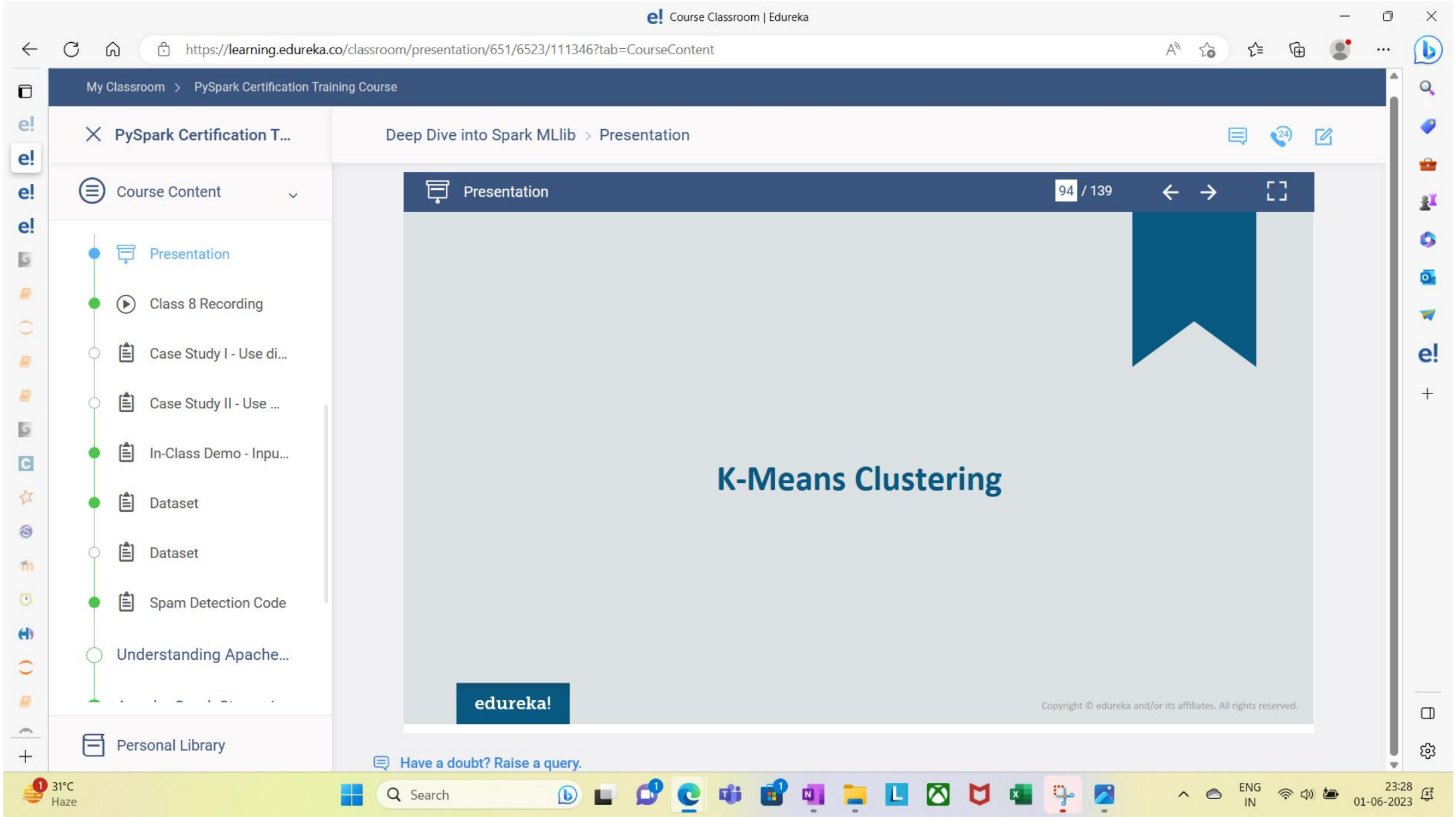
Hierarchical Clustering

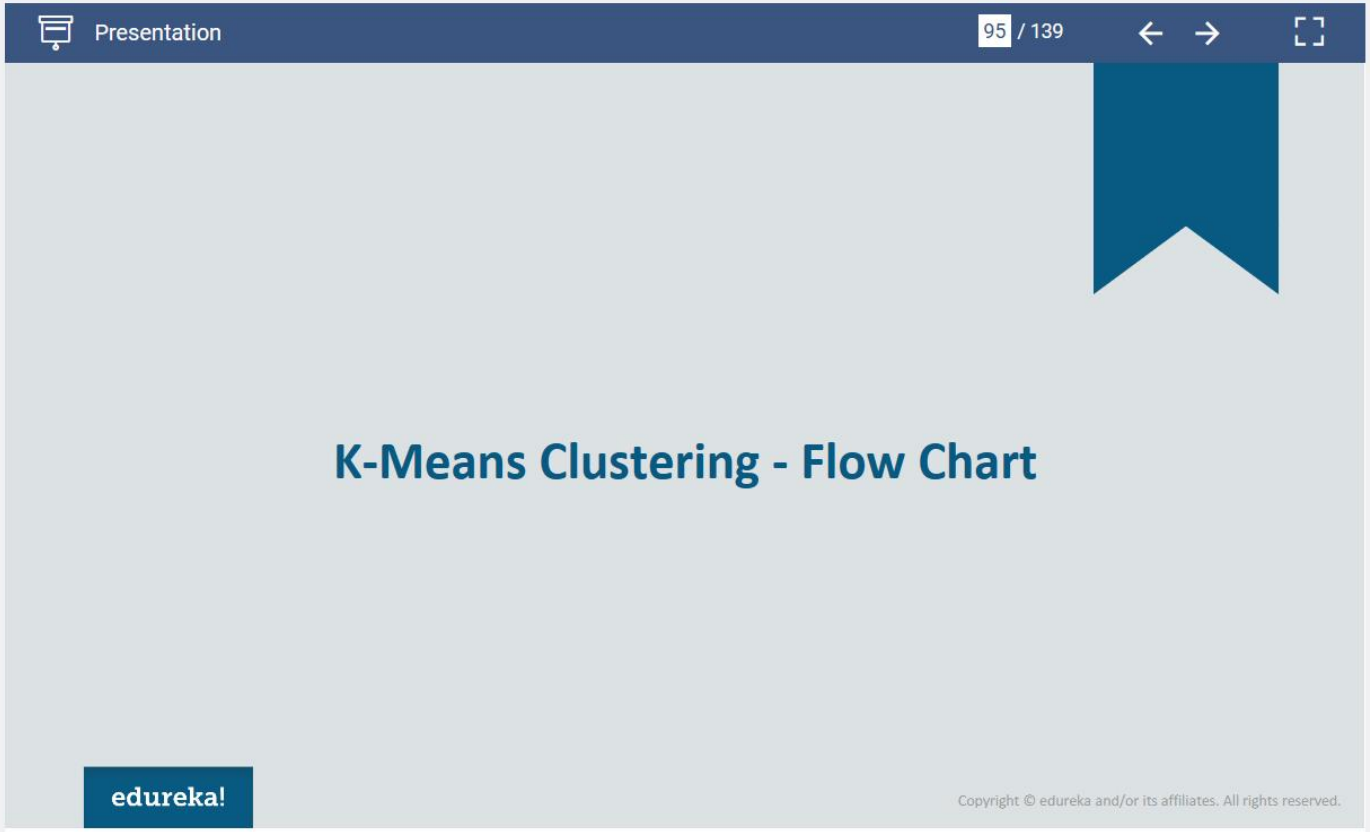
When two cluster have a parent-child relationship or a tree-like structure then it is Hierarchical clustering.



edureka!

Have a doubt? Raise a query.

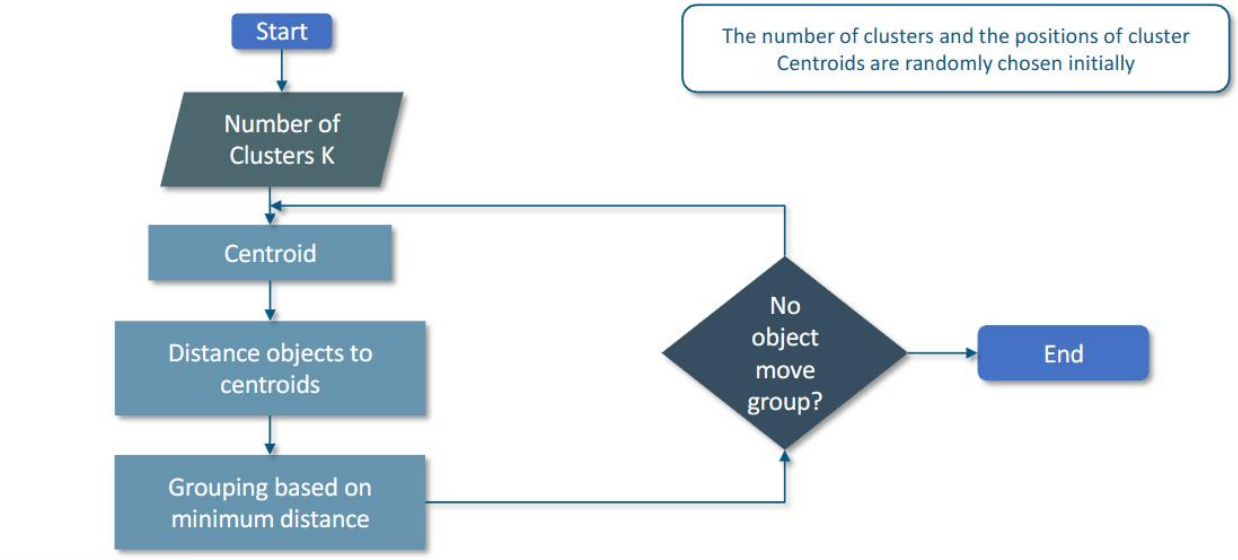

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
94 / 139
K-Means Clustering
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:28
01-06-2023



Have a doubt? Raise a query.

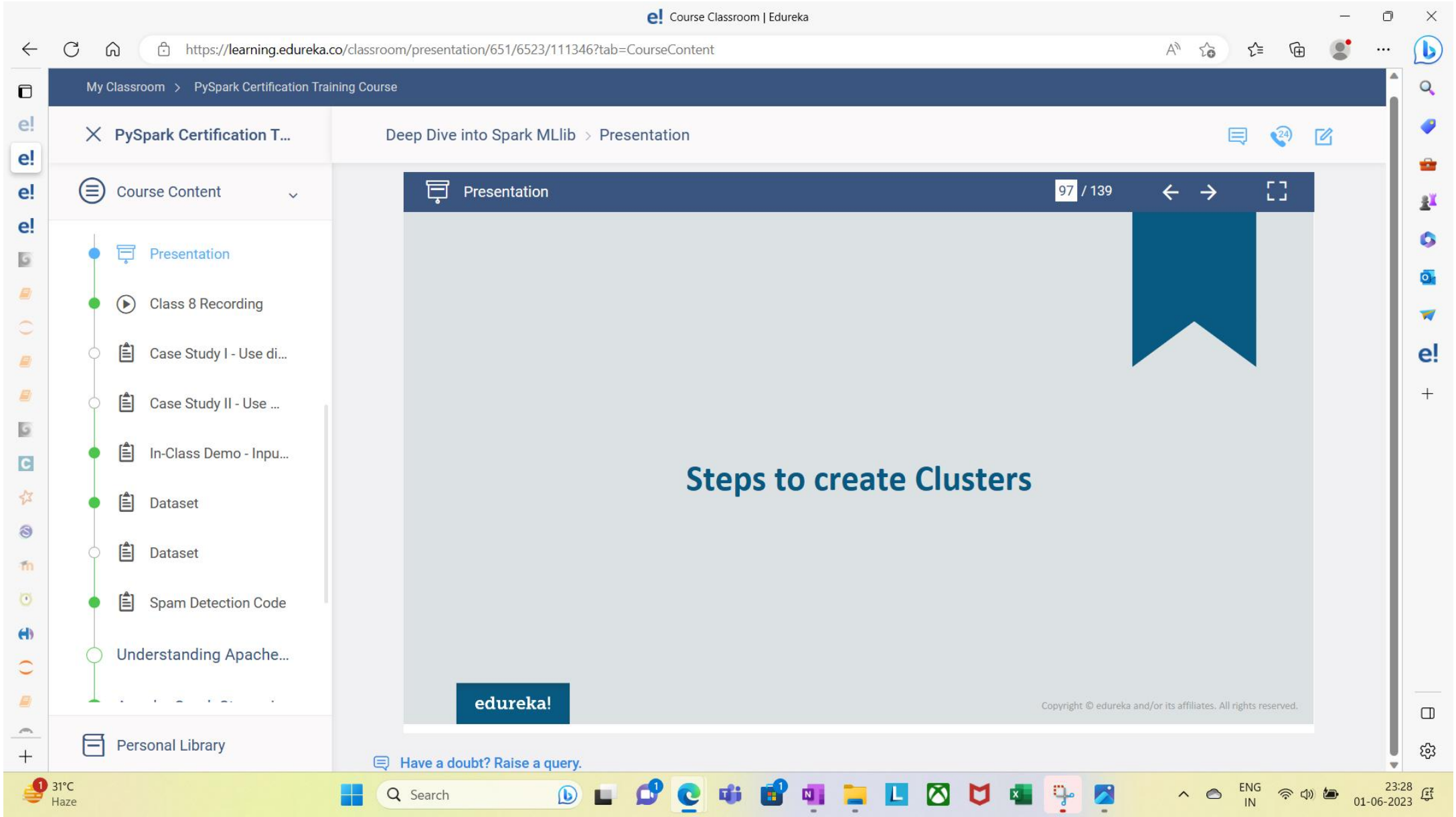
Presentation 96 / 139

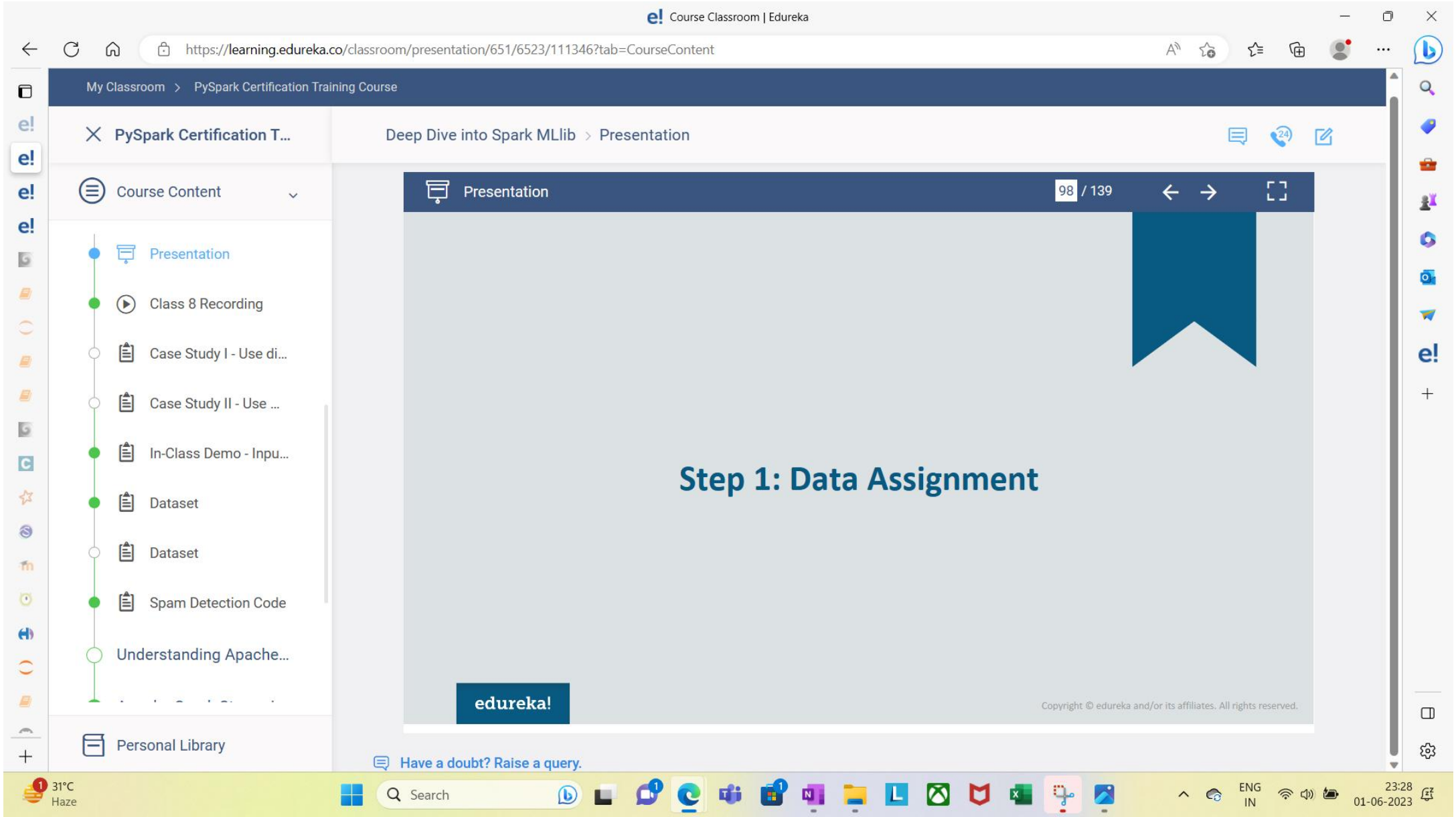
K-Means Clustering – Flow Chart



edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
97 / 139
Steps to create Clusters
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:28
01-06-2023


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
98 / 139
Step 1: Data Assignment
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:28
01-06-2023

Presentation 99 / 139

Data Assignment

- Cluster centroids are randomly placed in the plot. Each point is assigned to it's nearest centroid, based on the squared Euclidean distance
- More formally, if **ci** is the collection of centroids in set **C**, then each data point **x** is assigned to a cluster based on

$$\underset{c_i \in C}{\operatorname{argmin}}\ dist(c_i, x)^2$$

dist(.) is the Euclidean distance. Let the set of data point assignment for each **ith** cluster centroid be **si**

edureka!

Have a doubt? Raise a query.

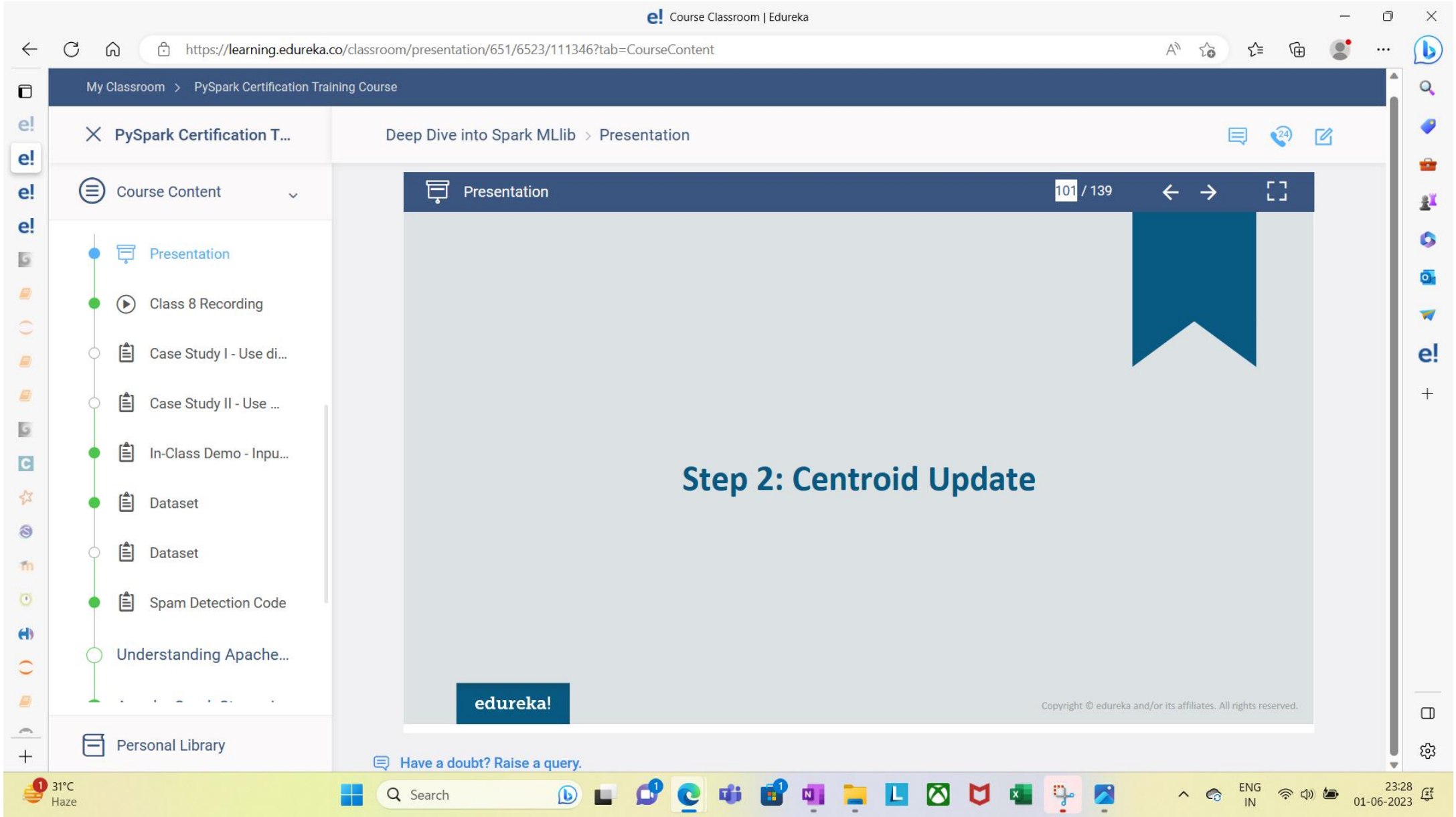
Euclidean Distance

dist(.) is the Euclidean distance. the **Euclidean distance** or **Euclidean metric** is the "ordinary" straight-line distance between two points in Euclidean space

$$dist(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

edureka!

Have a doubt? Raise a query.

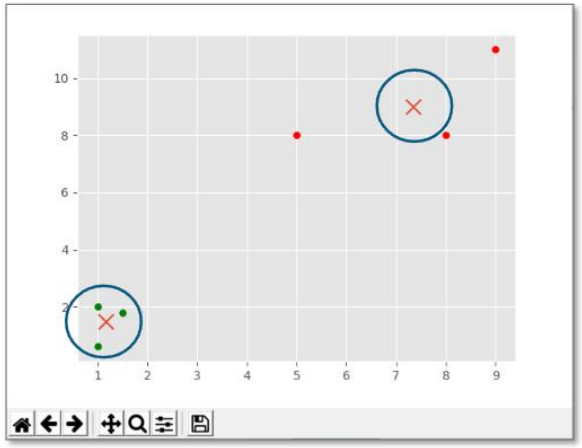

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
101 / 139
Step 2: Centroid Update
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:28
01-06-2023

Presentation 102 / 139

Centroid Update

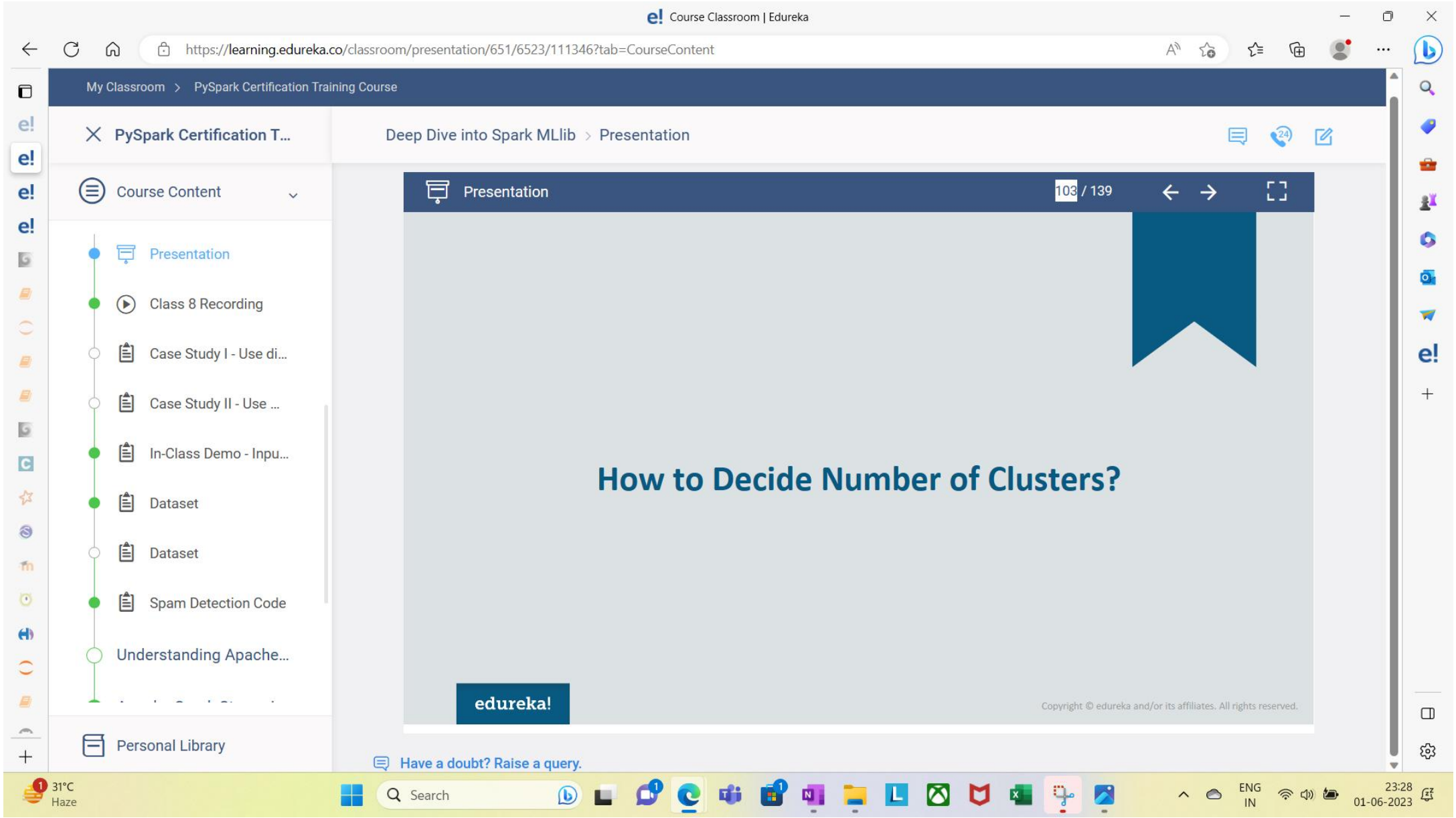
Centroids are recomputed

$$c_i = \frac{1}{|S_i|} \sum_{x_i \geq S_i} x_i$$



edureka!

Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
103 / 139
How to Decide Number of Clusters?
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:28
01-06-2023

Presentation 104 / 139

How to Decide Number of Clusters?

The Elbow Method :

First of all, compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid. Mathematically:

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist(x, c_i)^2$$

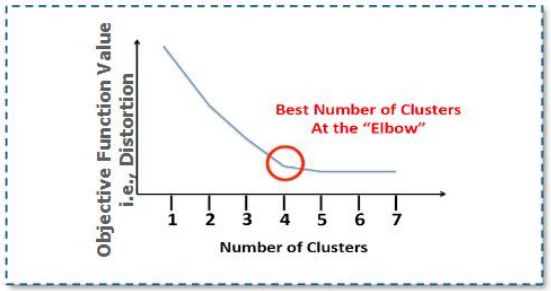
edureka!

Have a doubt? Raise a query.

Presentation 105 / 139

How to Decide Number of Clusters?

- If you plot k against the SSE, you will see that the error decreases as k gets larger, this is because when the number of clusters increases, they should be smaller, so distortion is also smaller.
- The idea of the elbow method is to choose the k at which the SSE decreases abruptly

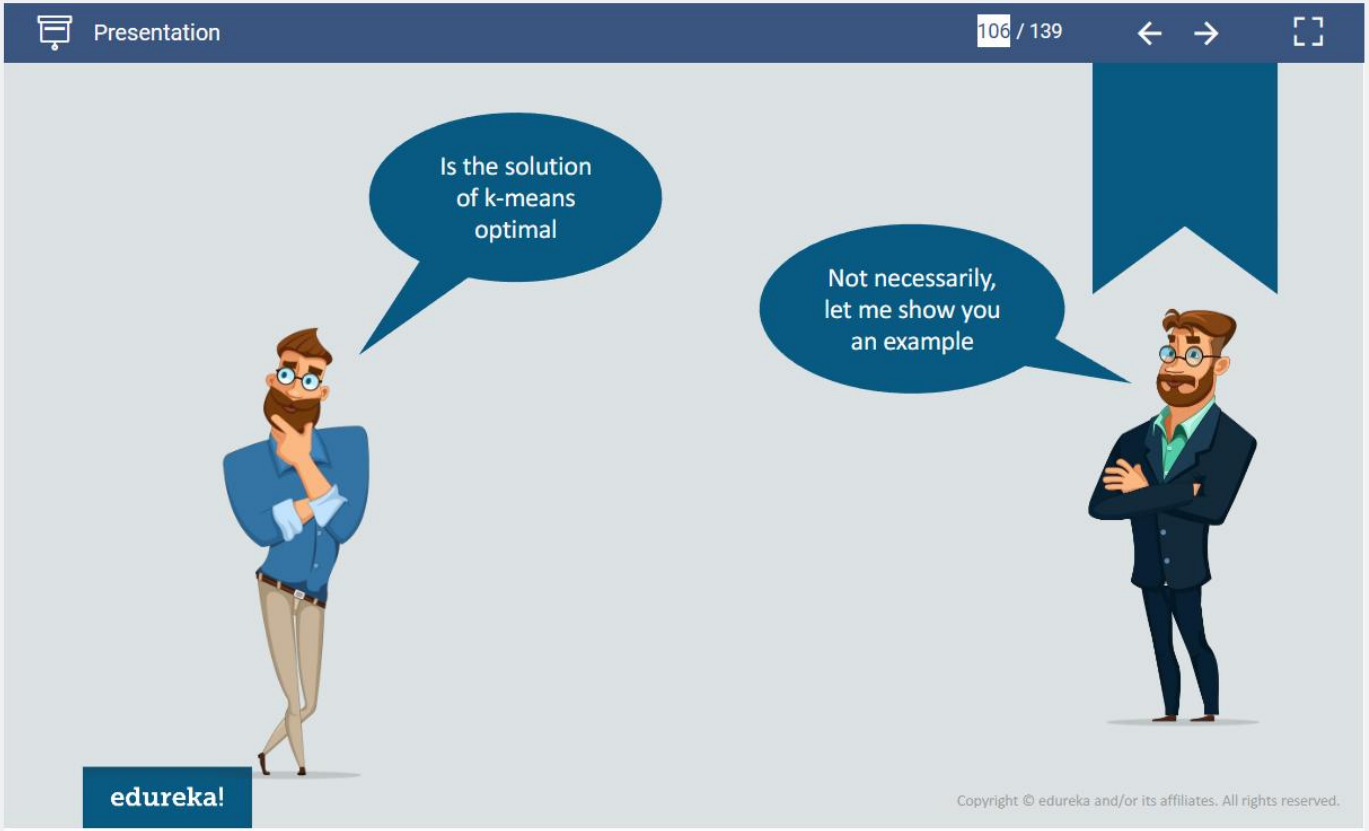


Elbow Plot

edureka!

Have a doubt? Raise a query.

Presentation 106 / 139

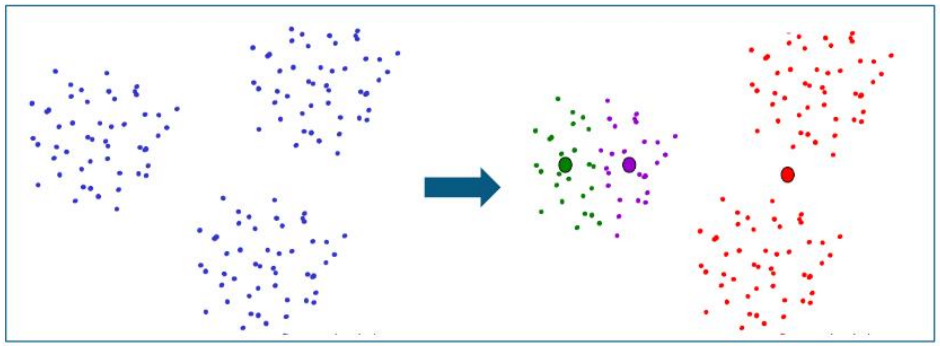


Have a doubt? Raise a query.

Presentation 107 / 139

Will We Find Optimal Solution?

Here we can see that we didn't get the clusters as we thought



But we can try to make it optimal, let's see how

edureka!

Have a doubt? Raise a query.

Presentation 108 / 139

How to Find Optimal Solution?

Idea 1: Careful about where we start

- Choose first center at random
- Choose second center that is far away from the first center
- Choose j^{th} center as far away as possible from the closest of centers 1 through (j-1)

Idea 2: Do many runs of K-means, each with different random starting point

edureka!

Have a doubt? Raise a query.

Presentation 109 / 139

Use Case 1: K-Means Clustering

edureka!

Have a doubt? Raise a query.

Presentation 110 / 139

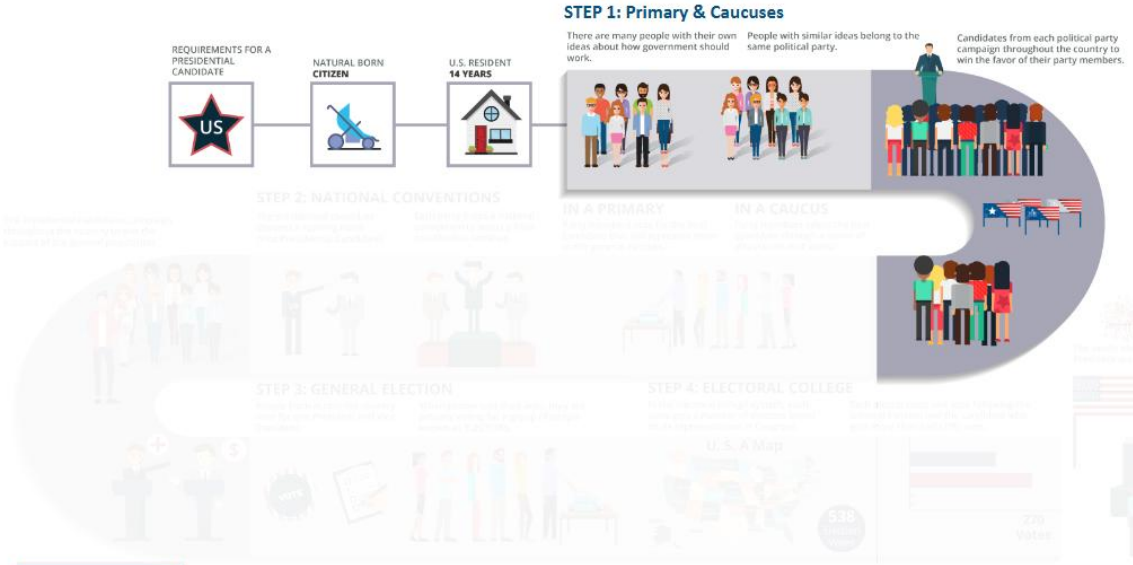
US Primary Election Analysis : USE CASE

edureka!

Have a doubt? Raise a query.

Presentation 111 / 139

US Election

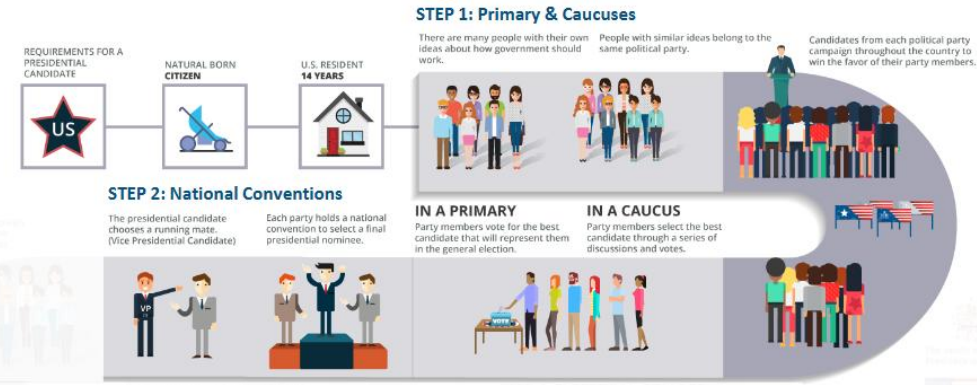


edureka!

Have a doubt? Raise a query.

Presentation 112 / 139

US Election

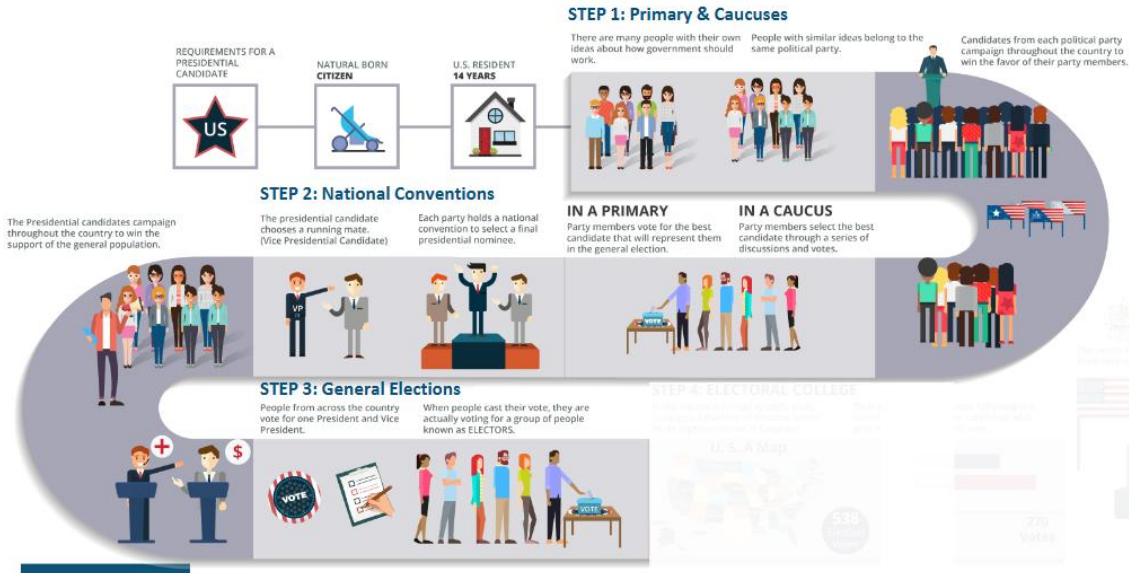


edureka!

Have a doubt? Raise a query.

Presentation 113 / 139

US Election

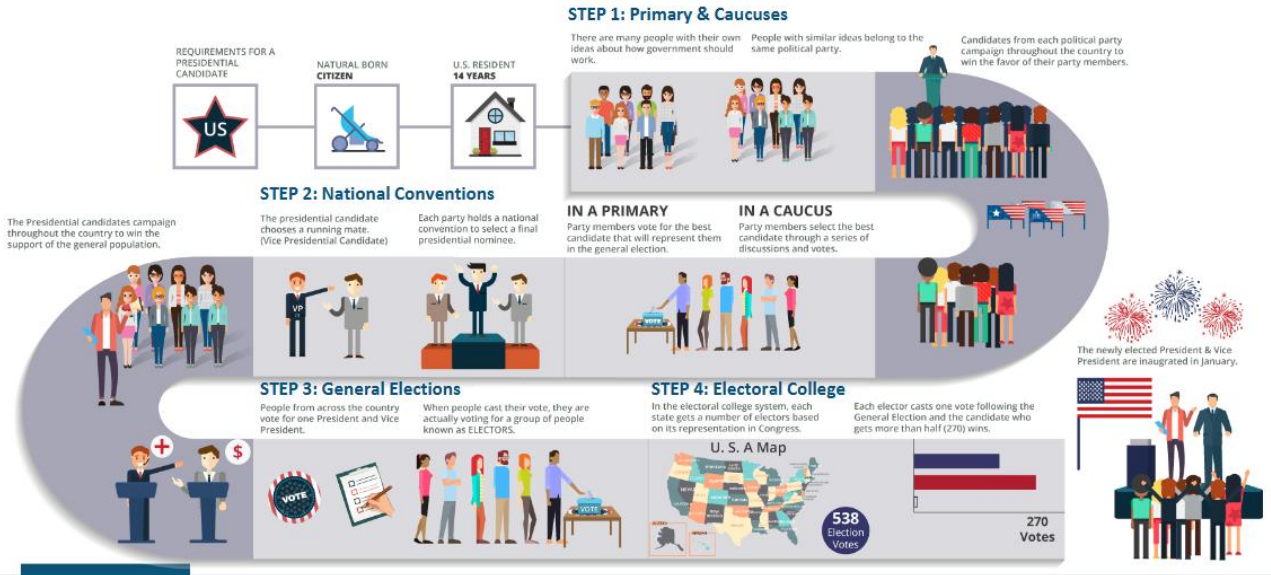


edureka!

Have a doubt? Raise a query.

Presentation 114 / 139

US Election



Have a doubt? Raise a query.

Presentation 115 / 139

US Primary Election

PROBLEM STATEMENT:

- In the US Primary Election 2016, **Hillary Clinton was nominated over Bernie Sanders from Democrats** and on the other hand, **Donald Trump was nominated from Republican Party** to contest for the presidential position.
- **As an analyst,** you have been tasked to **understand different factors** that **led to the winning of Hillary Clinton and Donald Trump** in the primary elections **based on demographic features** to plan their next initiatives and campaigns.

Republican *Democrat*

UNITED STATES OF AMERICA

edureka!

Have a doubt? Raise a query.

US Primary Election Dataset

Now as a data analyst you have 2 datasets available :

US Primary Election Data Set

state	state_abbreviation	county	fips	party	candidate	votes	fraction_votes
Alabama	AL	Autauga	1001	Democrat	Bernie Sanders	544	0.182
Alabama	AL	Autauga	1001	Democrat	Hillary Clinton	2387	0.8
Alabama	AL	Baldwin	1003	Democrat	Bernie Sanders	2694	0.329
Alabama	AL	Baldwin	1003	Democrat	Hillary Clinton	5290	0.647
Alabama	AL	Barbour	1005	Democrat	Bernie Sanders	222	0.078
Alabama	AL	Barbour	1005	Democrat	Hillary Clinton	2567	0.906
Alabama	AL	Bibb	1007	Democrat	Bernie Sanders	246	0.197
Alabama	AL	Bibb	1007	Democrat	Hillary Clinton	942	0.755
Alabama	AL	Blount	1009	Democrat	Bernie Sanders	395	0.386
Alabama	AL	Blount	1009	Democrat	Hillary Clinton	564	0.551
Alabama	AL	Bullock	1011	Democrat	Bernie Sanders	178	0.066
Alabama	AL	Bullock	1011	Democrat	Hillary Clinton	2451	0.913
Alabama	AL	Butler	1013	Democrat	Bernie Sanders	156	0.065

US Demographic Features (County-wise) Data Set

fips	area_name	state_abbreviation	PST045214	PST040210	PST120214	POP010210	AGE135214	AGE295214	AGE775214	SEX255214
0	United States		318857056	308758105	3.3	306745538	6.2	23.1	14.5	50.8
1000	Alabama		4849377	4780127	1.4	4779736	6.1	22.8	15.3	51.5
1001	Autauga County	AL	55395	54571	1.5	54571	6	25.2	13.8	51.4
1003	Baldwin County	AL	200111	182265	9.8	182265	5.6	22.2	18.7	51.2
1005	Barbour County	AL	26887	27457	-2.1	27457	5.7	21.2	16.5	46.8
1007	Bibb County	AL	22506	22919	-1.8	22915	5.3	21	14.8	45.9
1009	Blount County	AL	57719	57322	0.7	57322	6.1	23.6	17	50.5
1011	Bullock County	AL	10764	10915	-1.4	10914	6.3	21.4	14.9	45.8
1013	Butler County	AL	20256	20946	-3.1	20947	6.1	23.6	18	53.6
1015	Calhoun County	AL	115916	118586	-2.3	118572	5.7	22.2	16	51.8
1017	Chambers County	AL	34076	34170	-0.6	34215	5.9	21.4	18.3	52.3
1019	Cherokee County	AL	25037	25986	0.2	25989	4.8	20.4	20.9	50.2
1021	Chilton County	AL	43931	43631	0.7	43843	6.4	24.2	15.2	50.8
1023	Choctaw County	AL	13523	13858	-3.9	13859	4.9	20.6	20.8	52.5

edureka!

Have a doubt? Raise a query.

Presentation 117 / 139

US Primary Election Dataset

state	state_abbreviation	county	fips	party	candidate	votes	fraction_votes
Alabama	AL	Autauga	1001	Democrat	Bernie Sanders	544	0.182
Alabama	AL	Autauga	1001	Democrat	Hillary Clinton	2387	0.8
Alabama	AL	Baldwin	1003	Democrat	Bernie Sanders	2694	0.329
Alabama	AL	Baldwin	1003	Democrat	Hillary Clinton	5290	0.647
Alabama	AL	Barbour	1005	Democrat	Bernie Sanders	222	0.078
Alabama	AL	Barbour	1005	Democrat	Hillary Clinton	2567	0.906
Alabama	AL	Bibb	1007	Democrat	Bernie Sanders	246	0.197
Alabama	AL	Bibb	1007	Democrat	Hillary Clinton	942	0.755
Alabama	AL	Blount	1009	Democrat	Bernie Sanders	395	0.386
Alabama	AL	Blount	1009	Democrat	Hillary Clinton	564	0.551
Alabama	AL	Bullock	1011	Democrat	Bernie Sanders	178	0.066
Alabama	AL	Bullock	1011	Democrat	Hillary Clinton	2451	0.913
Alabama	AL	Butler	1013	Democrat	Bernie Sanders	156	0.065

state: List of US states
state_abbreviation: Abbreviation of each US state
county: List of counties in each US states
fips: FIPS county code is a Federal Information Processing Standards (FIPS) code which uniquely identifies counties

party: Different parties in US (i.e. Repulican & Democrat)
candidate: candidates in US primary election from different parties
votes: number of votes gained by a candidate
fraction_votes: total number of votes gained by a candidate/ total votes gained by the party

edureka!

Have a doubt? Raise a query.

Presentation 118 / 139

US County Demographic Features Dataset

fips	area_name	state_abbreviation	PST045214	PST040210	PST120214	POP010210	AGE135214	AGE295214	AGE775214	SEX255214
0	United States		318857056	308758105	3.3	308745538	6.2	23.1	14.5	50.8
1000	Alabama		4849377	4780127	1.4	4779736	6.1	22.8	15.3	51.5
1001	Autauga County	AL	55395	54571	1.5	54571	6	25.2	13.8	51.4
1003	Baldwin County	AL	200111	182265	9.8	182265	5.6	22.2	18.7	51.2
1005	Barbour County	AL	26887	27457	-2.1	27457	5.7	21.2	16.5	46.6
1007	Bibb County	AL	22506	22919	-1.8	22915	5.3	21	14.8	45.9
1009	Blount County	AL	57719	57322	0.7	57322	6.1	23.6	17	50.5
1011	Bullock County	AL	10764	10915	-1.4	10914	6.3	21.4	14.9	45.3
1013	Butler County	AL	20296	20946	-3.1	20947	6.1	23.6	18	53.6
1015	Calhoun County	AL	115916	118586	-2.3	118572	5.7	22.2	16	51.8
1017	Chambers County	AL	34076	34170	-0.3	34215	5.9	21.4	18.3	52.3
1019	Cherokee County	AL	26037	25986	0.2	25989	4.8	20.4	20.9	50.2
1021	Chilton County	AL	43931	43631	0.7	43643	6.4	24.2	15.2	50.8
1023	Choctaw County	AL	13323	13858	-3.9	13859	4.9	20.6	20.8	52.5

DETAILS:

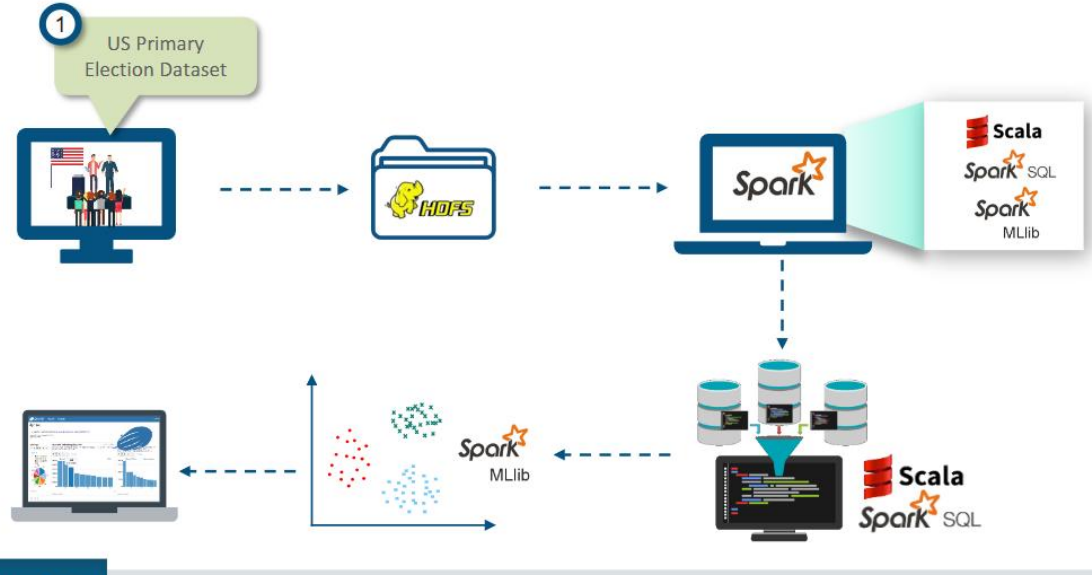
Population, 2014 estimate
Population, 2010 (April 1) estimates base
Population, percent change - April 1, 2010 to July 1, 2014
Population, 2010
Persons under 5 years, percent, 2014
Persons under 18 years, percent, 2014
Persons 65 years and over, percent, 2014
Female persons, percent, 2014
White alone, percent, 2014 ...

edureka!

Have a doubt? Raise a query.

Presentation 119 / 139

US Election Solution Strategy

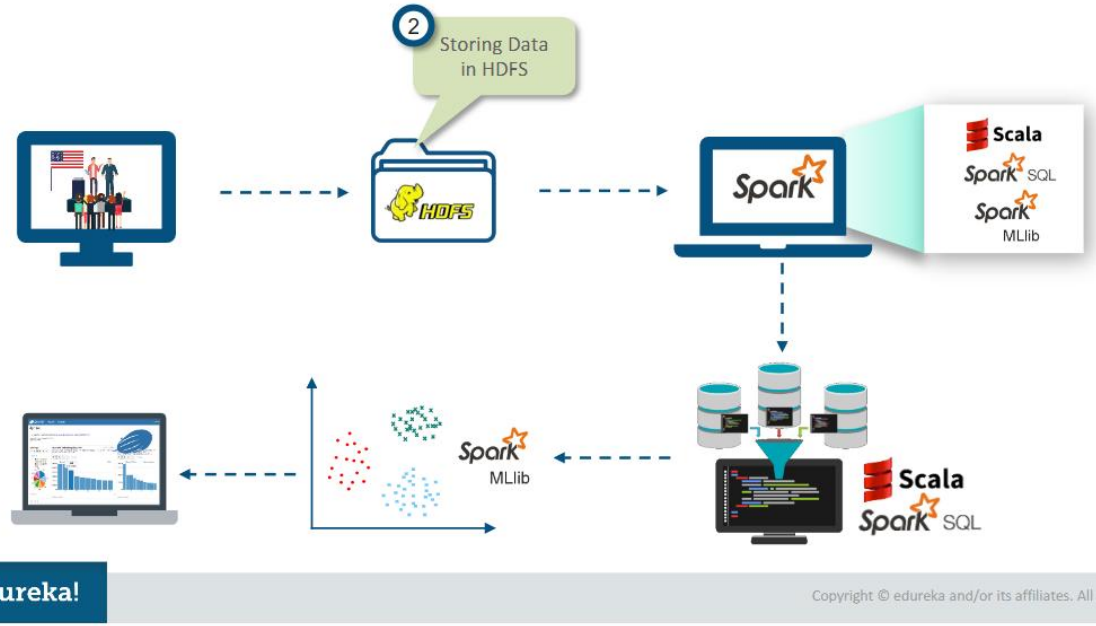


edureka!

Have a doubt? Raise a query.

Presentation 120 / 139

US Election Solution Strategy

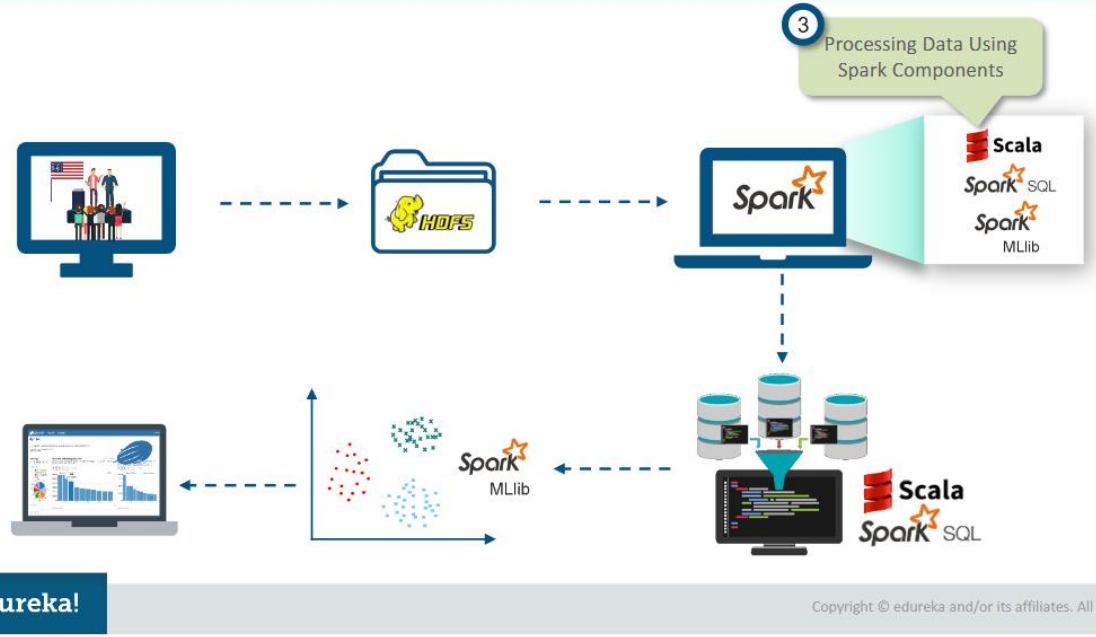


edureka!

Have a doubt? Raise a query.

Presentation 121 / 139

US Election Solution Strategy

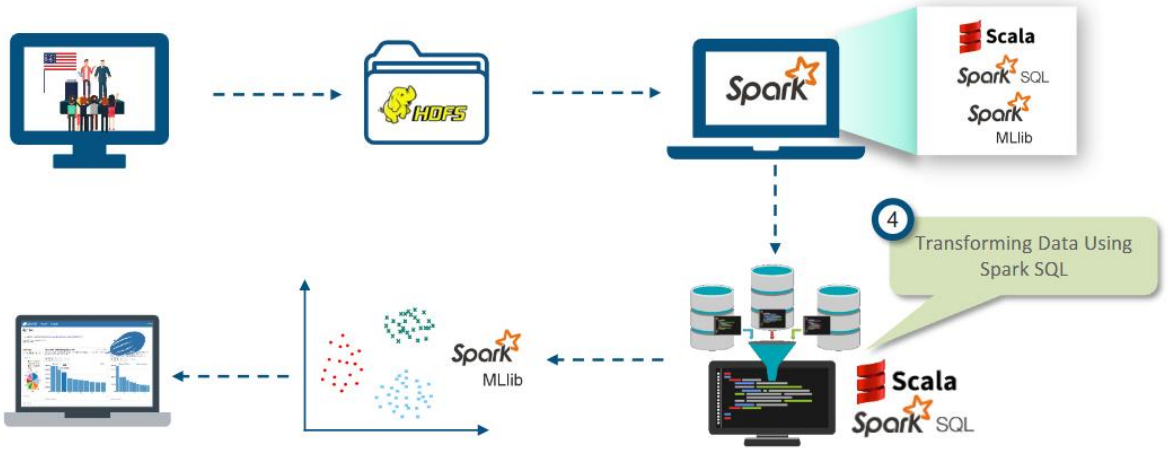


edureka!

Have a doubt? Raise a query.

Presentation 122 / 139

US Election Solution Strategy

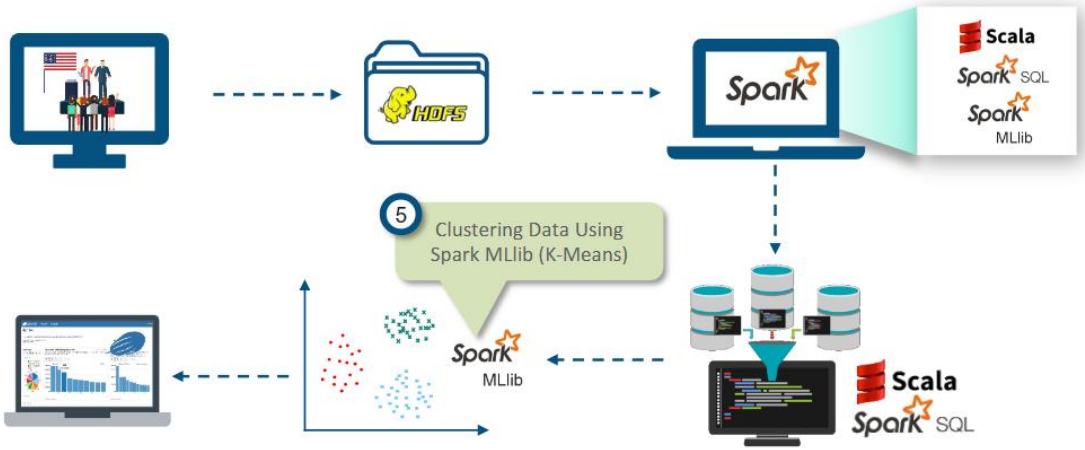


edureka!

Have a doubt? Raise a query.

Presentation 123 / 139

US Election Solution Strategy

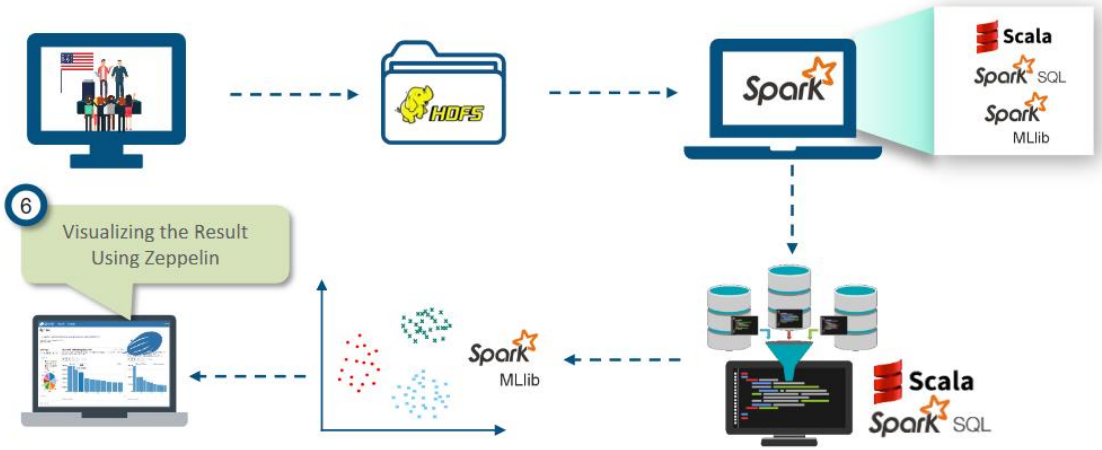


edureka!

Have a doubt? Raise a query.

Presentation 124 / 139

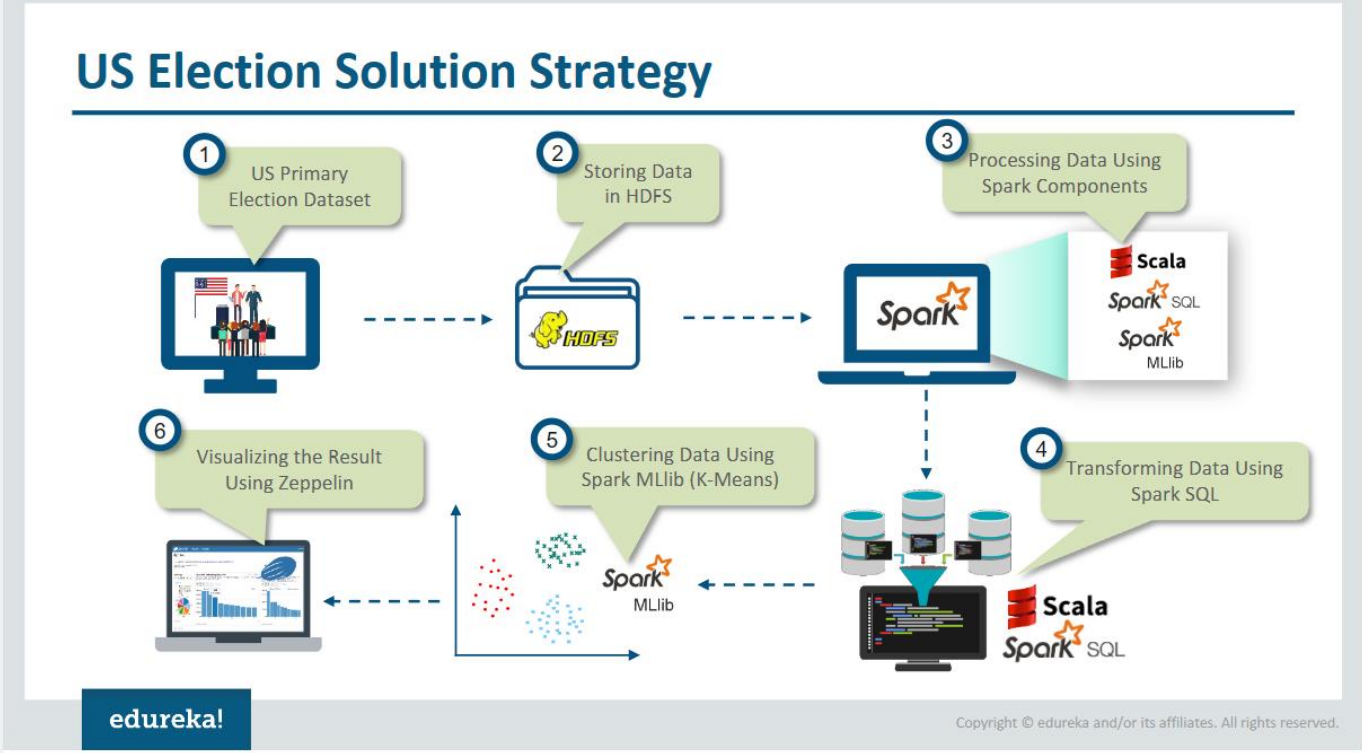
US Election Solution Strategy



edureka!

Have a doubt? Raise a query.

Presentation 125 / 139



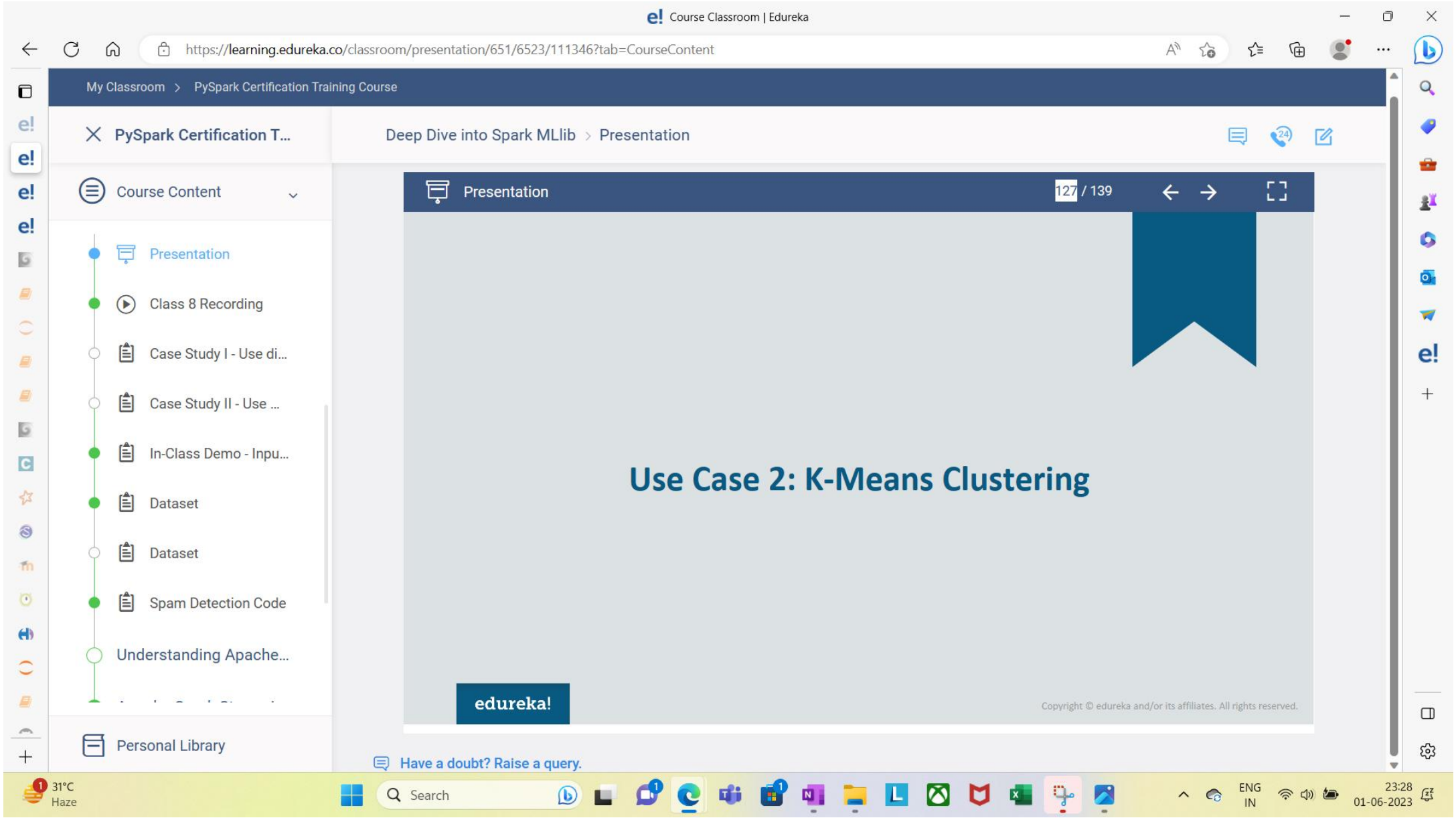
Have a doubt? Raise a query.

Presentation 126 / 139

Validation of Result

edureka!

Have a doubt? Raise a query.

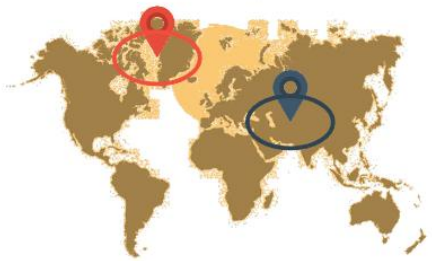

Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
127 / 139
Use Case 2: K-Means Clustering
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:28
01-06-2023

Presentation 128 / 139

Problem Statement

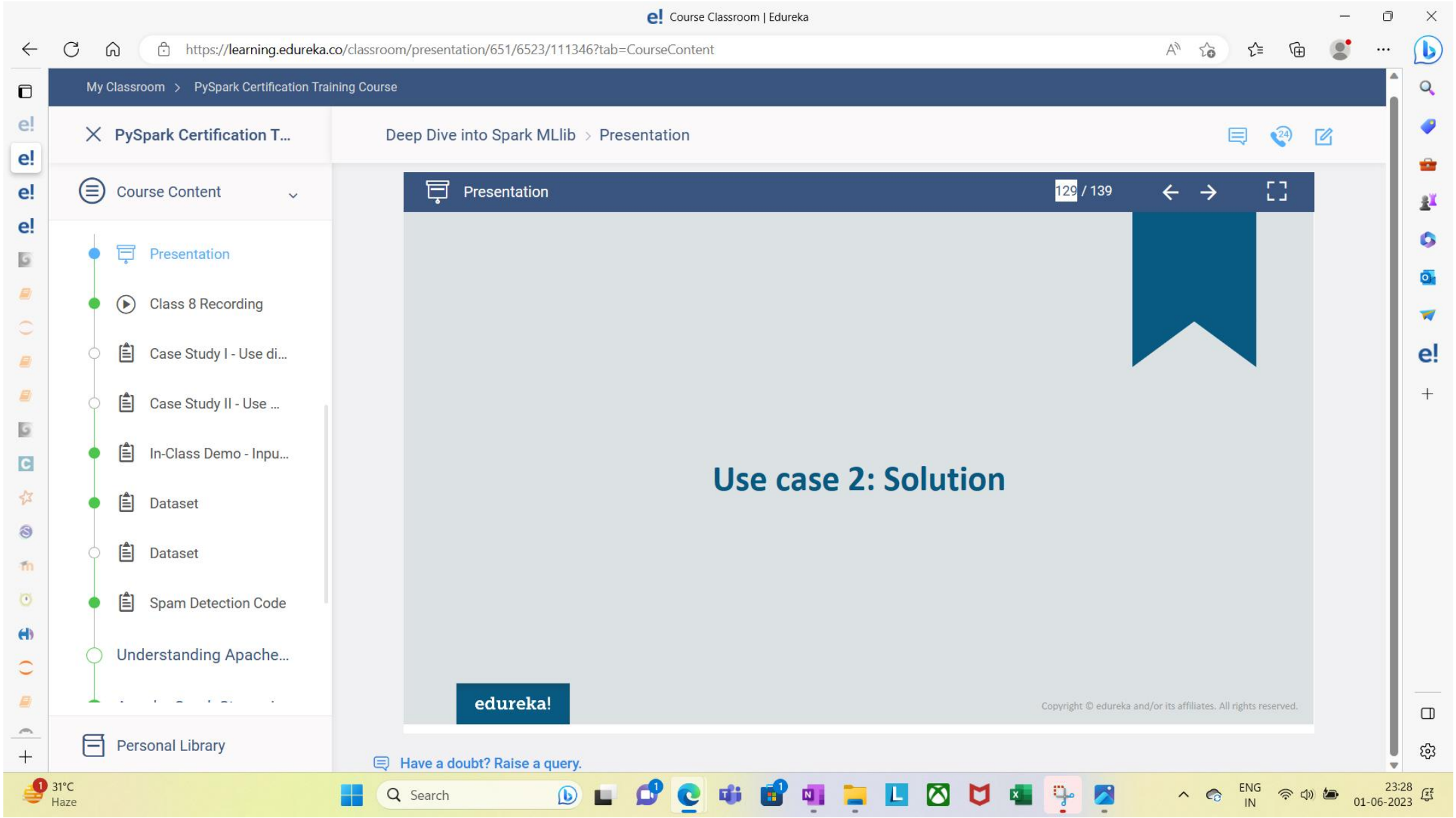
John is a Geologist. He has a set of Latitudes and Longitudes. He wants to make 2 groups of the locations, so that he can find out which places are near to each other

John



edureka!

Have a doubt? Raise a query.


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
129 / 139
Use case 2: Solution
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:28
01-06-2023

Presentation 130 / 139

Use Case 2 Solution: K-Means Clustering

```
In [1]: from pyspark import SparkContext
        from pyspark.sql import SQLContext

In [2]: sc = SparkContext()
        sqlContext = SQLContext(sc)

In [3]: df = sqlContext.createDataFrame([[0, 33.3, -17.5],
                                         [1, 40.4, -20.5],
                                         [2, 28., -23.9],
                                         [3, 29.5, -19.0],
                                         [4, 32.8, -18.84]
                                        ],["other","lat", "long"])

In [4]: df.show()
+-----+----+------+
|other| lat|  long|
+-----+----+------+
|    0|33.3| -17.5|
|    1|40.4| -20.5|
|    2|28.0| -23.9|
|    3|29.5| -19.0|
|    4|32.8|-18.84|
+-----+----+------+
```

Dataframe of the Latitudes and Longitudes is created

Used to display the data present in the Dataframe

edureka!

Have a doubt? Raise a query.

Use Case 2 Solution: K-Means Clustering

```
In [5]: from pyspark.ml.feature import VectorAssembler

In [6]: vecAssembler = VectorAssembler(inputCols=["lat", "long"], outputCol="features")
        new_df = vecAssembler.transform(df)
        new_df.show()
```

```
+-----+----+------+-------------+
|other| lat|  long|     features|
+-----+----+------+-------------+
|    0|33.3| -17.5| [33.3,-17.5]|
|    1|40.4| -20.5| [40.4,-20.5]|
|    2|28.0| -23.9| [28.0,-23.9]|
|    3|29.5| -19.0| [29.5,-19.0]|
|    4|32.8|-18.84|[32.8,-18.84]|
+-----+----+------+-------------+
```

```
In [7]: from pyspark.ml.clustering import KMeans

In [8]: kmeans = KMeans(k=2, seed=1)  # 2 clusters here
        model = kmeans.fit(new_df.select('features'))
```

VectorAssembler class is imported from pyspark.ml.feature library

transform() function is used for Scaling and modifying the features

Number of Clusters are defined (K=2). fit() method is used to normalize each standard feature of Dataframe

edureka!

Have a doubt? Raise a query.

Presentation 132 / 139

Use Case 2 Solution: K-Means Clustering

```
In [9]: transformed = model.transform(new_df)
        transformed.show()
```

```
+-----+----+------+-------------+----------+
|other| lat|  long|     features|prediction|
+-----+----+------+-------------+----------+
|    0|33.3| -17.5| [33.3,-17.5]|         0|
|    1|40.4| -20.5| [40.4,-20.5]|         1|
|    2|28.0| -23.9| [28.0,-23.9]|         0|
|    3|29.5| -19.0| [29.5,-19.0]|         0|
|    4|32.8|-18.84|[32.8,-18.84]|         0|
+-----+----+------+-------------+----------+
```

model.transform() method is used to make predictions. It will divide the locations in two different clusters

```
In [10]: df.select('lat', 'long').rdd.collect()
Out[10]: [Row(lat=33.3, long=-17.5),
          Row(lat=40.4, long=-20.5),
          Row(lat=28.0, long=-23.9),
          Row(lat=29.5, long=-19.0),
          Row(lat=32.8, long=-18.84)]
```

df.select() method is used to select the respective colunns from the dataframe

edureka!

Have a doubt? Raise a query.

Presentation 133 / 139

Use Case 2 Solution: K-Means Clustering

```
In [11]: df.select('lat', 'long').rdd.map(lambda x: (x[0], x[1])).collect()
Out[11]: [(33.3, -17.5), (40.4, -20.5), (28.0, -23.9), (29.5, -19.0), (32.8, -18.84)]

In [12]: from pyspark.mllib.clustering import KMeans, KMeansModel

         rdd = df.select('lat', 'long').rdd.map(lambda x: (x[0], x[1]))
         clusters = KMeans.train(rdd, 2, maxIterations=10, initializationMode="random")

In [13]: clusters.centers
Out[13]: [array([ 34.  , -18.96]), array([ 28. , -23.9])]
```

This will print the centroids of the 2 Clusters, in which the locations are divided

edureka!

Have a doubt? Raise a query.

Presentation 134 / 139

Pros and Cons of K-Means Clustering

Pros:

- Simple, understandable
- Items automatically assigned to clusters

Cons:

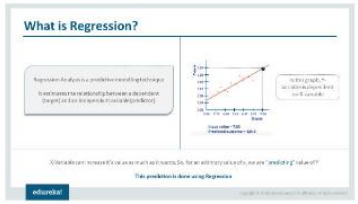
- Must define number of clusters
- All items forced into clusters
- Unable to handle noisy data and outliers

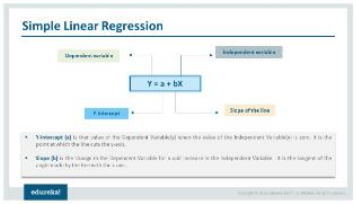
edureka!

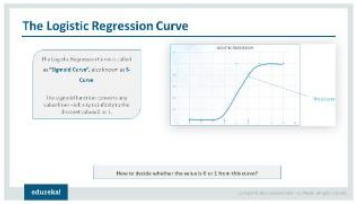
Have a doubt? Raise a query.

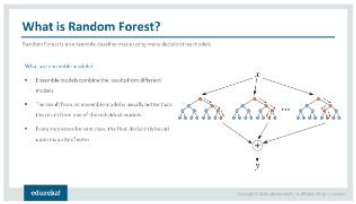
Presentation 135 / 139

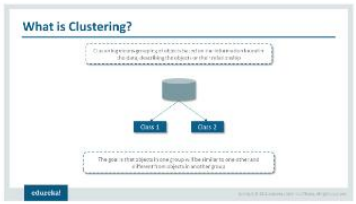
Summary





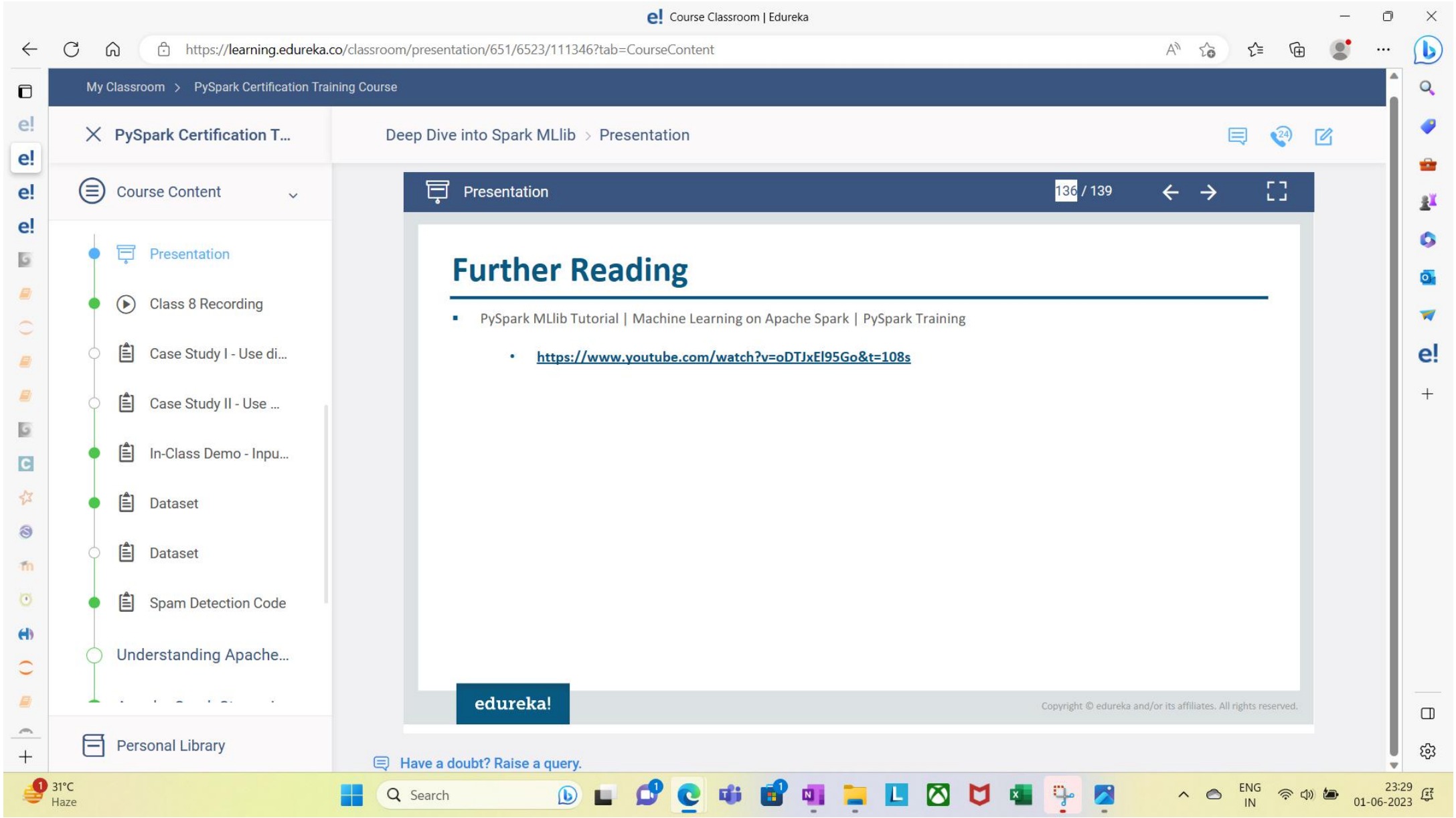






edureka!

Have a doubt? Raise a query.


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6523/111346?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Deep Dive into Spark MLlib > Presentation
Course Content
Presentation
Class 8 Recording
Case Study I - Use di...
Case Study II - Use ...
In-Class Demo - Inpu...
Dataset
Dataset
Spam Detection Code
Understanding Apache...
Personal Library
Presentation
136 / 139
Further Reading
PySpark MLlib Tutorial | Machine Learning on Apache Spark | PySpark Training
https://www.youtube.com/watch?v=oDTJxEl95Go&t=108s
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
Have a doubt? Raise a query.
31°C
Haze
Search
ENG
IN
23:29
01-06-2023