Presentation 3 / 78

COURSE OUTLINE

MODULE 07

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

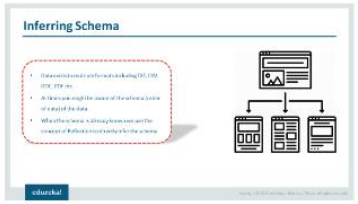
Apache Spark Streaming - Processing Multiple Batches

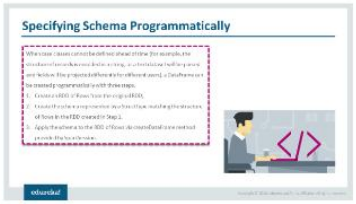
Apache Spark Streaming - Data Sources

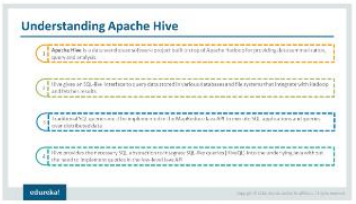
Implementing an End-to-End Project

Presentation 4 / 78

Recap





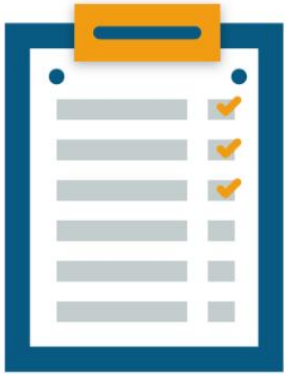


edureka!

Presentation 5 / 78

Topics

- What is Machine Learning
- Features of Machine Learning
- Applications of Machine Learning
- Steps of Machine Learning
- Types of Machine Learning
- Mlib in Spark



edureka!

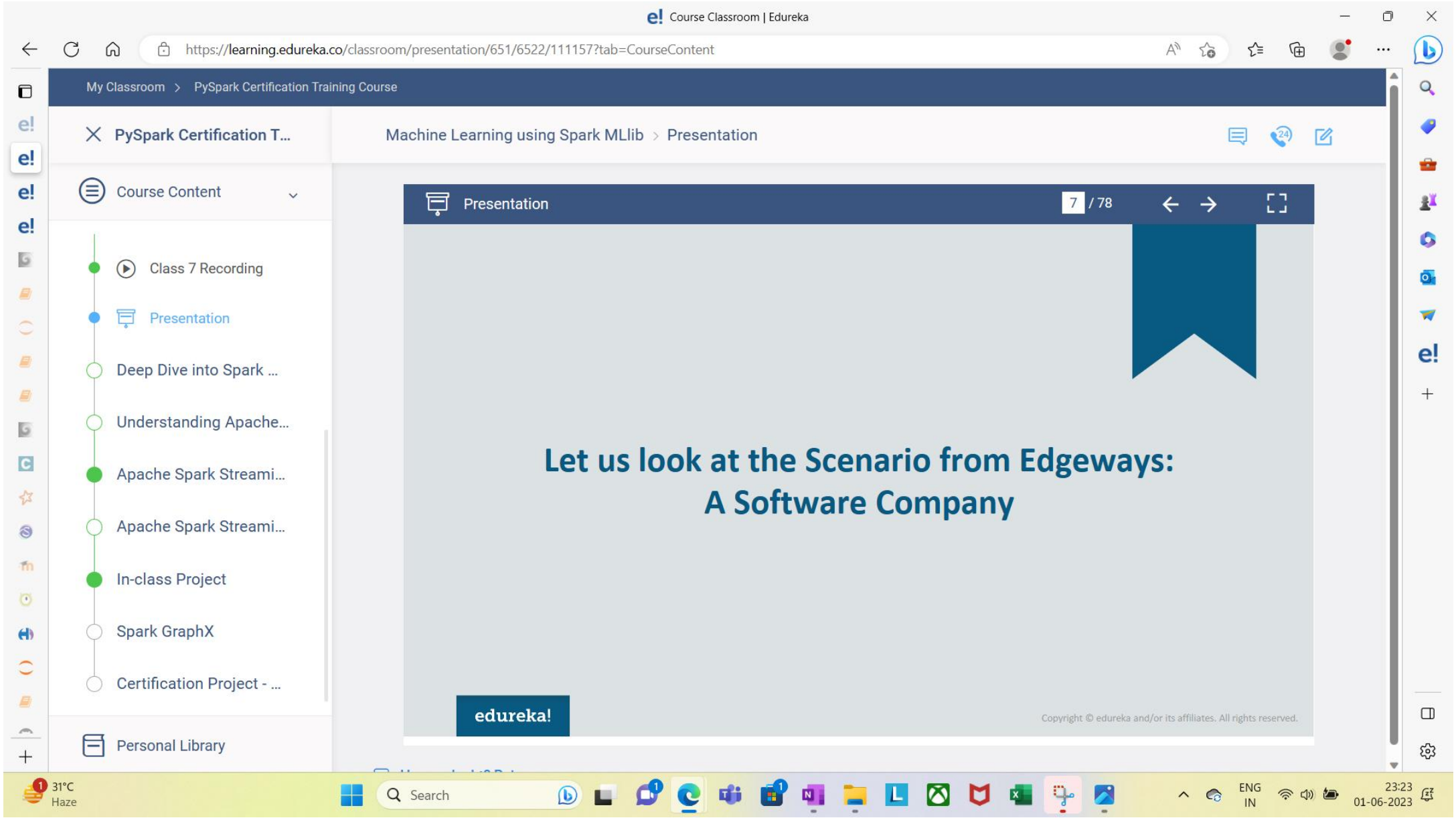
Presentation 6 / 78

Objectives

After completing this module, you should be able to:

- Describe What is Machine Learning and Where It is Used
- Describe Different Types of Machine Learning Techniques
- Understand Face Detection : USE CASE
- Describe MLlib
- Explain Features of MLlib and MLlib Tools
- Get an overview of various ML algorithms supported by MLlib

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
7 / 78
Let us look at the Scenario from Edgeways:
A Software Company
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:23
01-06-2023

Presentation 8 / 78

Scenario of Edgeways

A professional working in **Edgeways, a software company** checks inbox for new mails or updates and becomes frustrated after seeing unnecessary mails

Hey Dear,
Congratulations!!!!
I am pleased to inform you that you have won an amount of 100K in a lottery ticket. You can send your "account details" by replying to same mail.
Best Regards
Friend

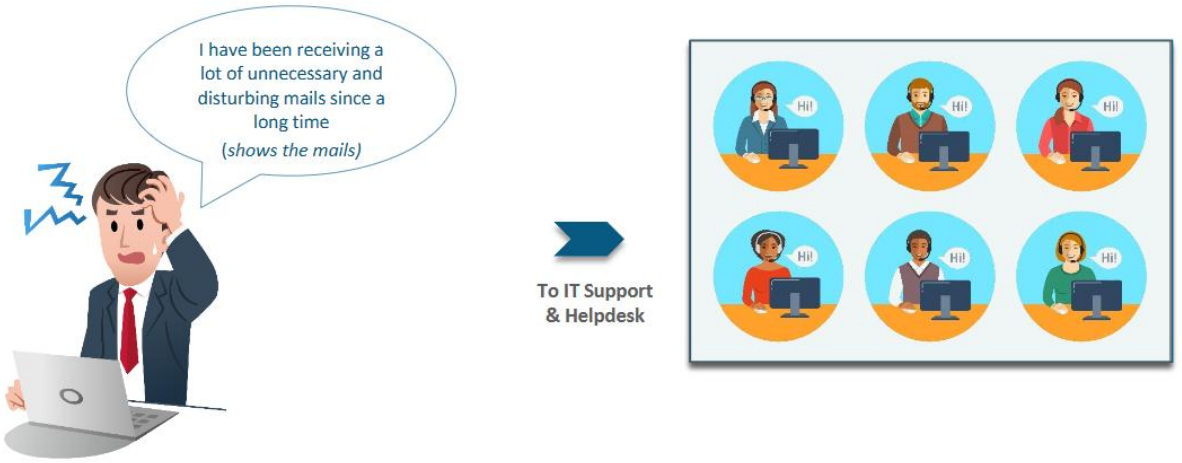
Hey Dear,
Surprises are waiting at your door!!!!
Reply to the same mail and see the magic.
Best Regards
Anonymous

many more

edureka!

Presentation 9 / 78

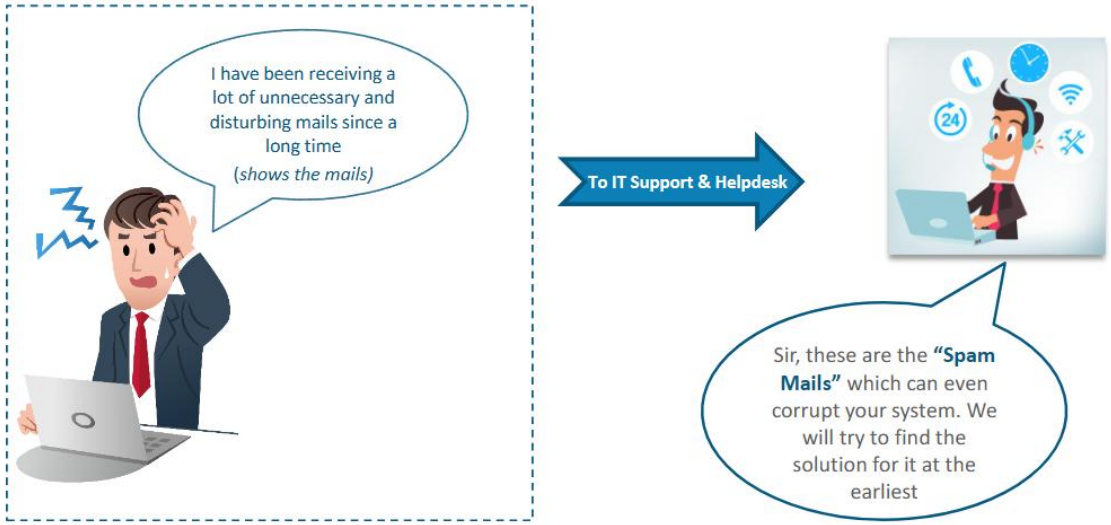
Edgeways: Challenge Faced



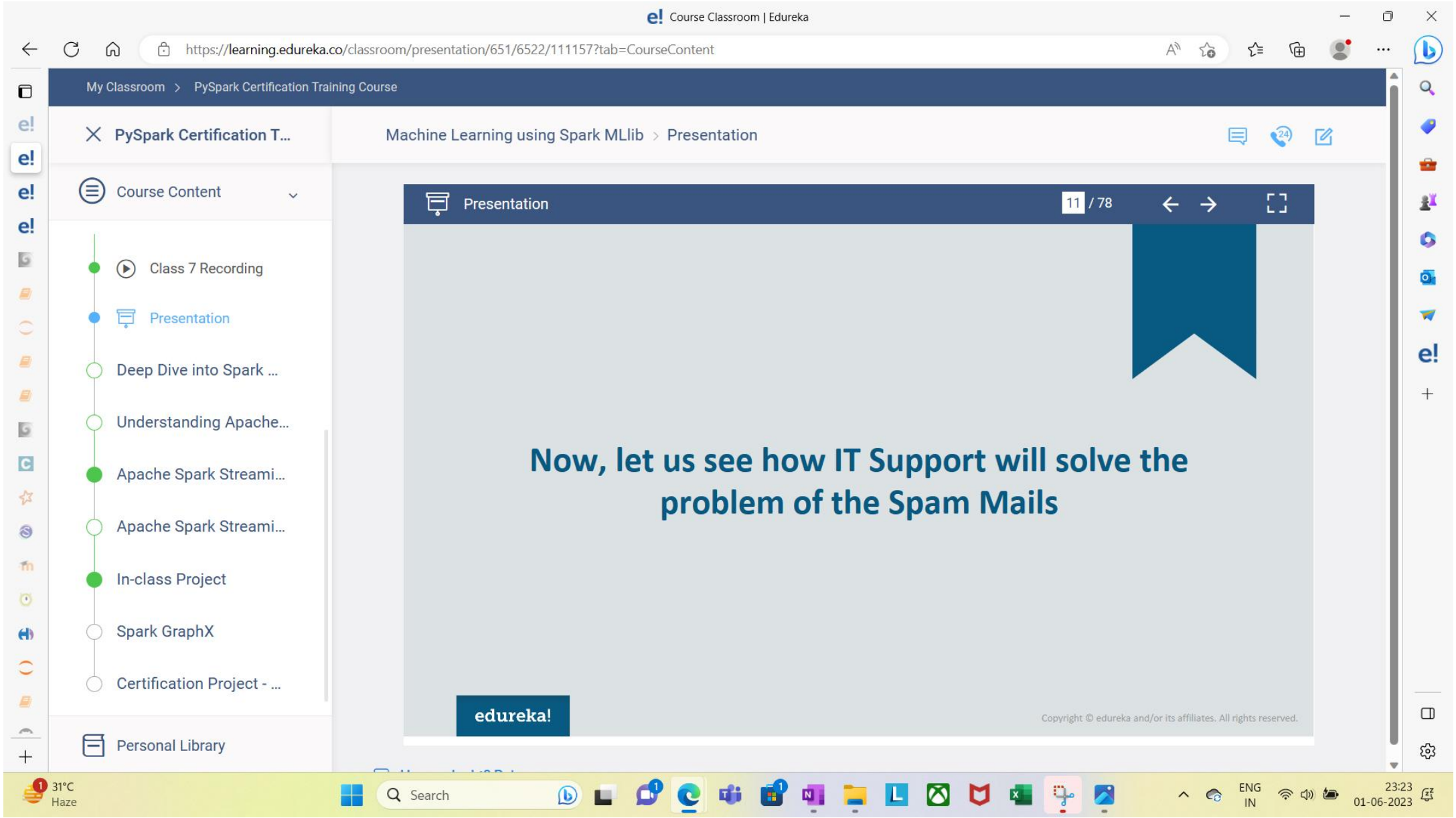
edureka!

Presentation 10 / 78

Edgeways: Challenge Faced



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
11 / 78
Now, let us see how IT Support will solve the problem of the Spam Mails
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:23
01-06-2023

Presentation 12 / 78

Probable Solution

Spam mails can be detected by manually setting a filter on some words

For Example,
If there are words like **lottery**, then mark the mail as a spam mail

edureka!

Presentation 13 / 78

Probable Solution

edureka!

Presentation 14 / 78

How can we solve this problem?

Who will help me to solve this problem?

edureka!

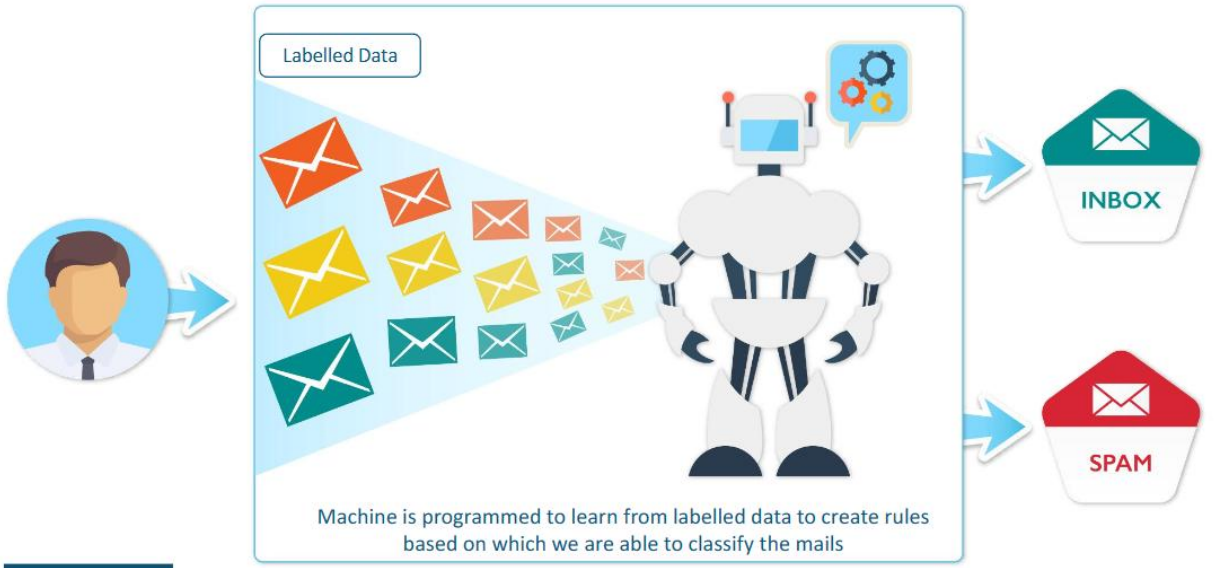
Presentation 15 / 78

How can we solve this problem?

edureka!

Presentation 16 / 78

Solution



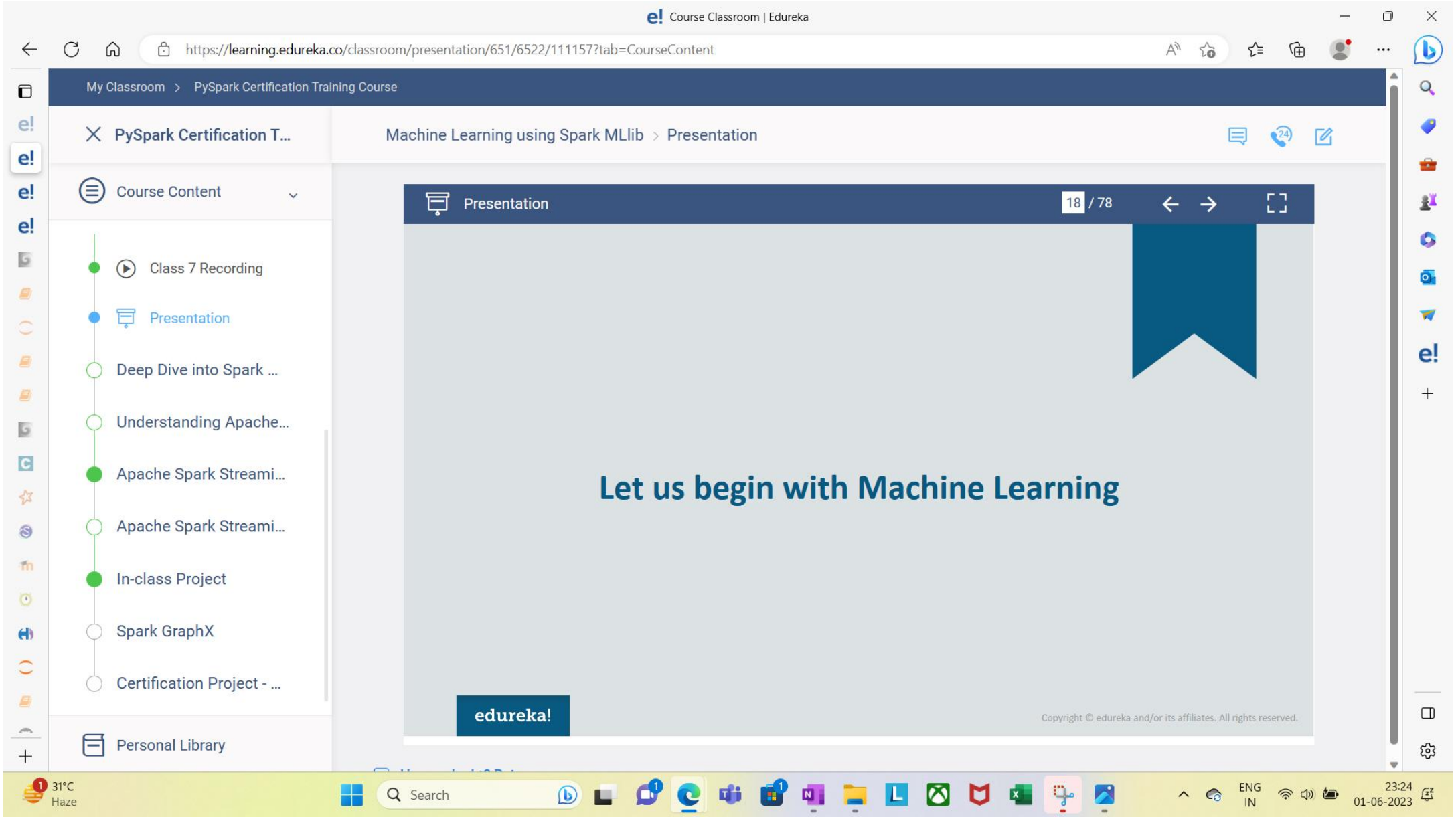
edureka!

Presentation 17 / 78

Solution

To get the required solution, Machine Learning is used. Let us understand what is Machine Learning

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
18 / 78
Let us begin with Machine Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 19 / 78

What is Machine Learning?

"**Machine learning** is a *method* of *data analysis* that *automates analytical model building*. Using algorithms that *iteratively learn from data*, machine learning allows computers to *find hidden insights without being explicitly programmed* where to look"

Customers	Last Purchase	Products in Cart	Recent Searches
Customer 1	Shoes	Socks, Sneakers	Wedges, Block Heels
Customer 2	Laptop	Laptop Charger, Laptop Bag	Mouse
Customer 3	Mobile Phone	Mobile Charger, Tempered Glass	Back Cover, HandHold
Customer 4	T-Shirt	Shorts, Jeans	Leggings, Kurti
Customer 5	Hand Bag	Sling Bag, Wallets	Clutches
Customer 6	SunGlasses	Cover, Specs	Camera, Handycam
Customer 7	Grocery	Glass, Bowl	Spoons, Cupset
Customer 8	Accessories	Ear-Ring, Rings	Nosepin,
Customer 9	Make-Up Kit	EyeLiner, Lipstic	Kajal, LipGloss
Customer 10	Wrist Watch	Mi-Band, Bracelet	Wall Clocks

Training Data

edureka!

Presentation 20 / 78

What is Machine Learning?

"**Machine learning** is a *method* of *data analysis* that *automates analytical model building*. Using algorithms that *iteratively learn from data*, machine learning allows computers to *find hidden insights without being explicitly programmed* where to look"

Customers	Last Purchase	Products in Cart	Recent Searches
Customer 1	Shoes	Socks, Sneakers	Wedges, Block Heels
Customer 2	Laptop	Laptop Charger, Laptop Bag	Mouse
Customer 3	Mobile Phone	Mobile Charger, Tempered Glass	Back Cover, HandHold
Customer 4	T-Shirt	Shorts, Jeans	Leggings, Kurti
Customer 5	Hand Bag	Sling Bag, Wallets	Clutches
Customer 6	SunGlasses	Cover, Specs	Camera, Handycam
Customer 7	Grocery	Glass, Bowl	Spoons, Cupset
Customer 8	Accessories	Ear-Ring, Rings	Nosepin,
Customer 9	Make-Up Kit	EyeLiner, Lipstic	Kajal, LipGloss
Customer 10	Wrist Watch	Mi-Band, Bracelet	Wall Clocks

Customers	Last Purchase	Products in Cart
Customer 1	Shoes	Socks, Sneakers
Customer 2	Laptop	Laptop Charger, Laptop Bag
Customer 3	Mobile Phone	Mobile Charger, Tempered Glass
Customer 4	T-Shirt	Shorts, Jeans
Customer 5	Hand Bag	Sling Bag, Wallets
Customer 6	SunGlasses	Cover, Specs
Customer 7	Grocery	Glass, Bowl
Customer 8	Accessories	Ear-Ring, Rings

Training Data

Accuracy

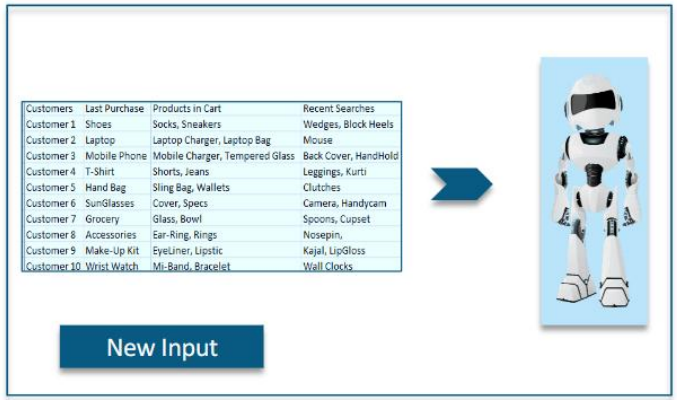
Testing Data

edureka!

Presentation 21 / 78

What is Machine Learning?

"**Machine learning** is a *method* of *data analysis* that *automates analytical model building*. Using algorithms that *iteratively learn from data*, machine learning allows computers to *find hidden insights without being explicitly programmed* where to look"



edureka!

Presentation 22 / 78

Features of Machine Learning

01

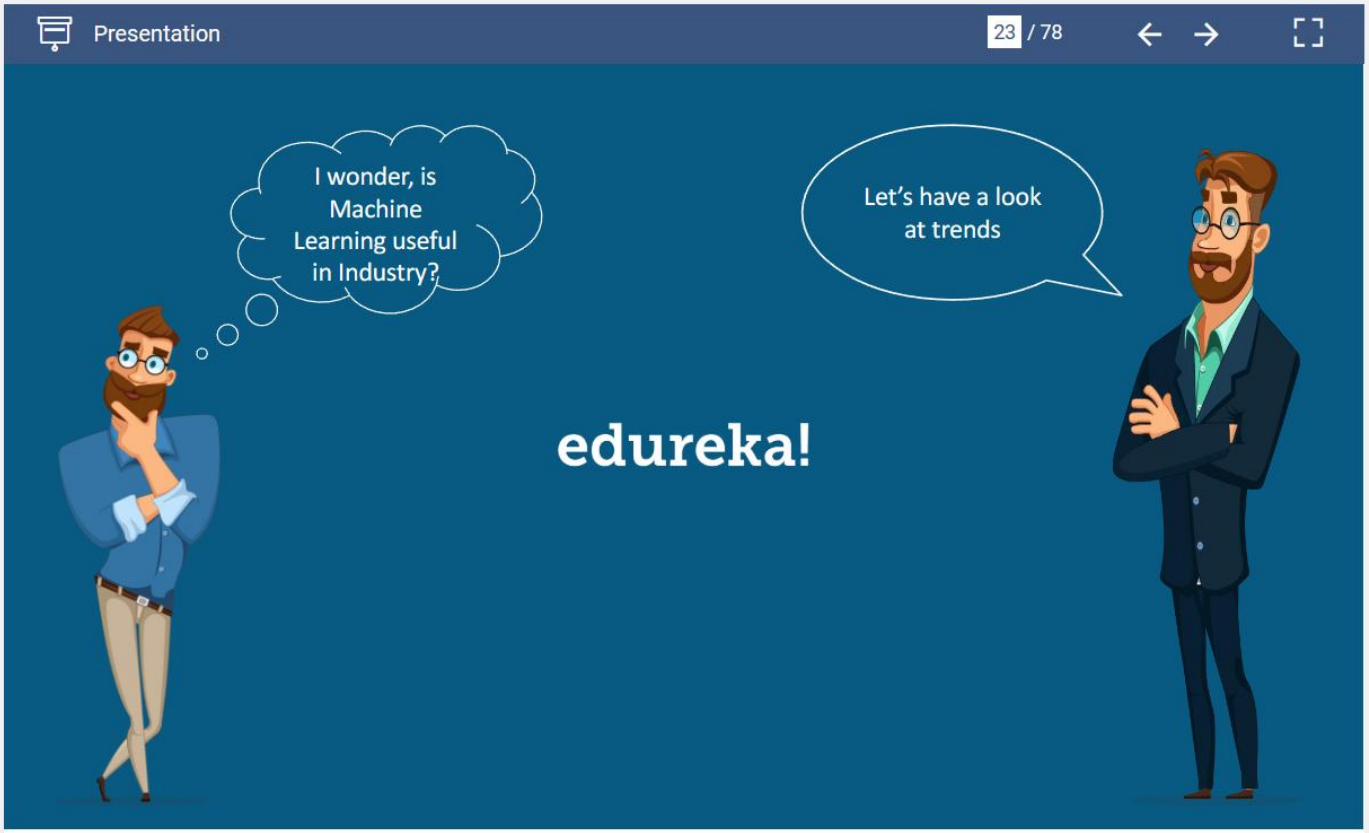
It uses the data to *detect patterns* in a dataset and *adjust program actions accordingly*

It *focuses on* the *development of computer programs* that can teach themselves to *grow and change* when *exposed* to *new data*

02

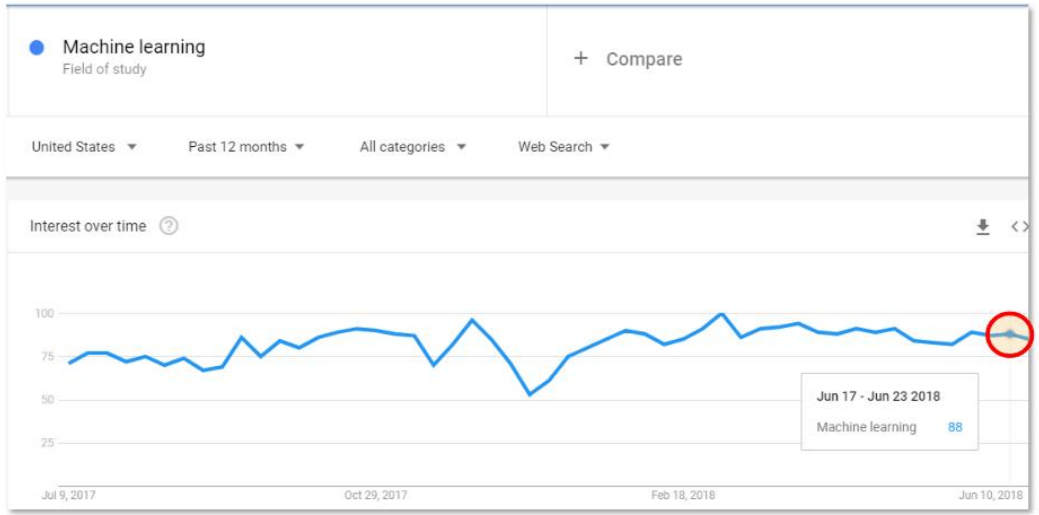
edureka!

Presentation 23 / 78

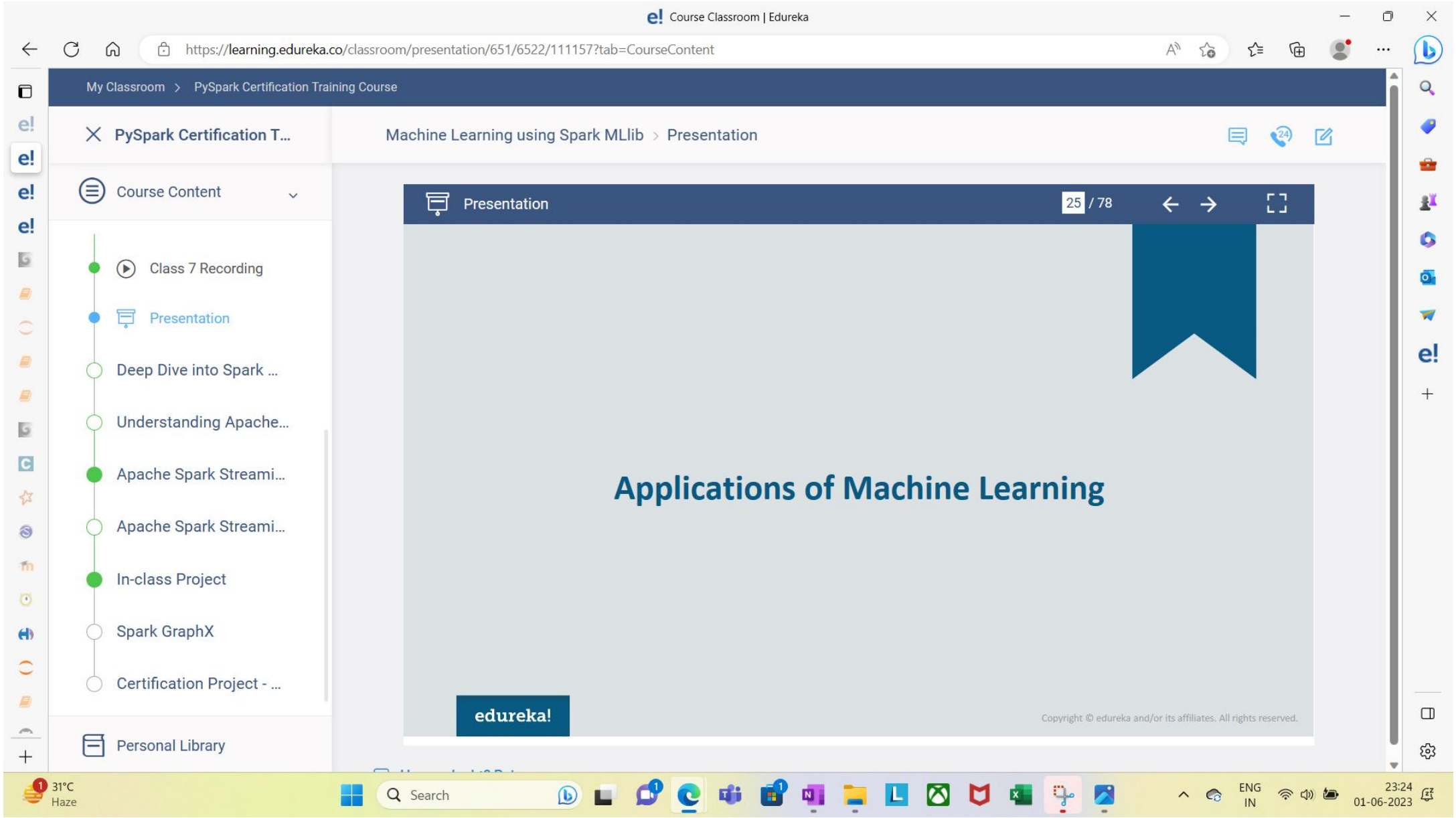


Presentation 24 / 78

Market Trend: Machine Learning



edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
25 / 78
Applications of Machine Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 26 / 78

Machine Learning Applications

Siri

Apple claims that the software adapts to user's individual preferences overtime and personalizes results

Marketing and Sales

This ability to capture data, analyse it and use it to personalize a shopping experience (or implement a marketing campaign) is the future of retail

edureka!

Presentation 27 / 78

Machine Learning Applications

Healthcare

Machine Learning involves the use of wearable devices and sensors to assess a patient's health in real time

Financial Services

Financial Industry use machine Learning technology to identify insights in data and prevent fraud

edureka!

Presentation 28 / 78

Machine Learning Applications

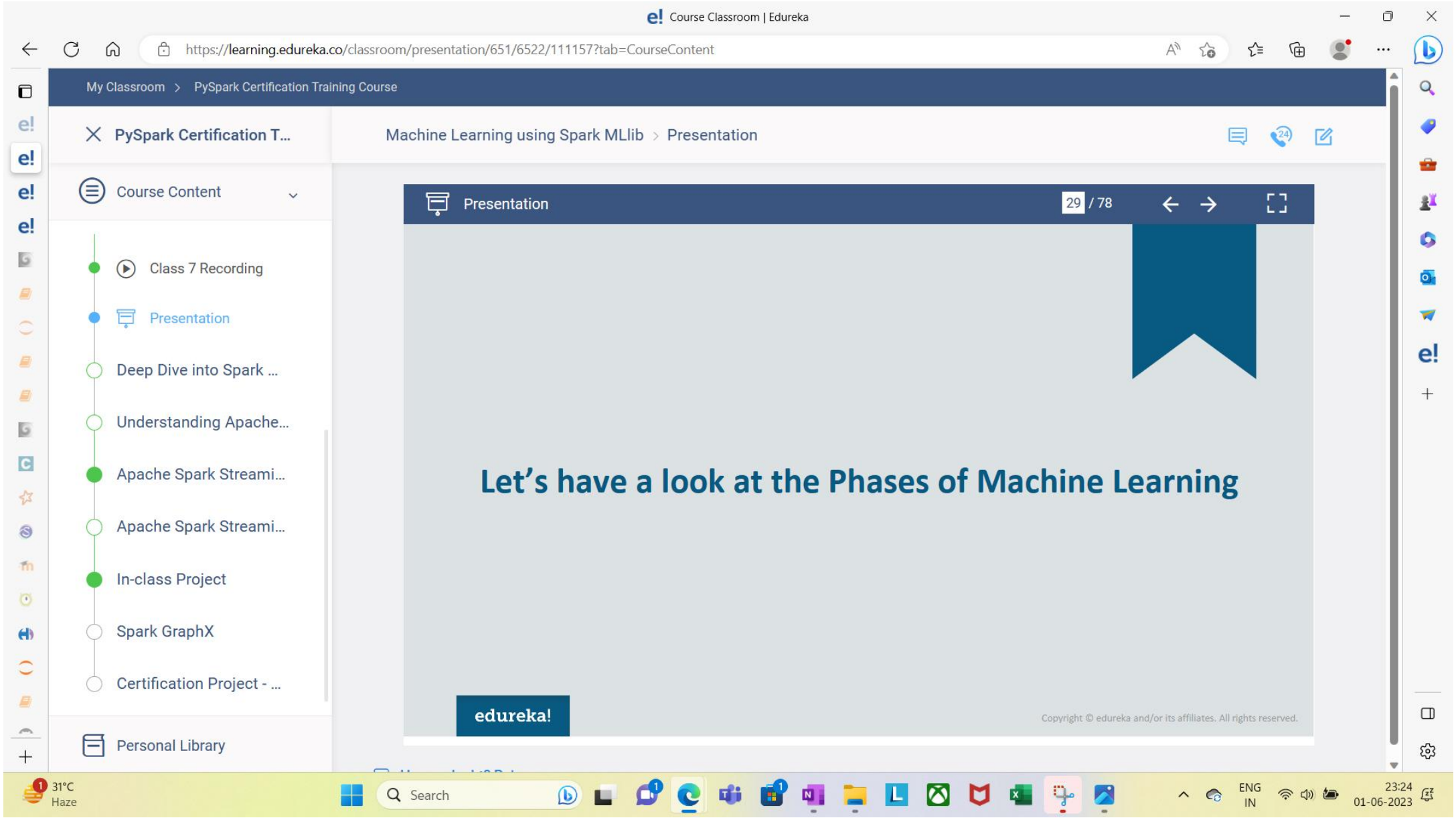
Biometrics

Biometrics use machine learning to identify individual credentials and prevent banking fraud

Transportation

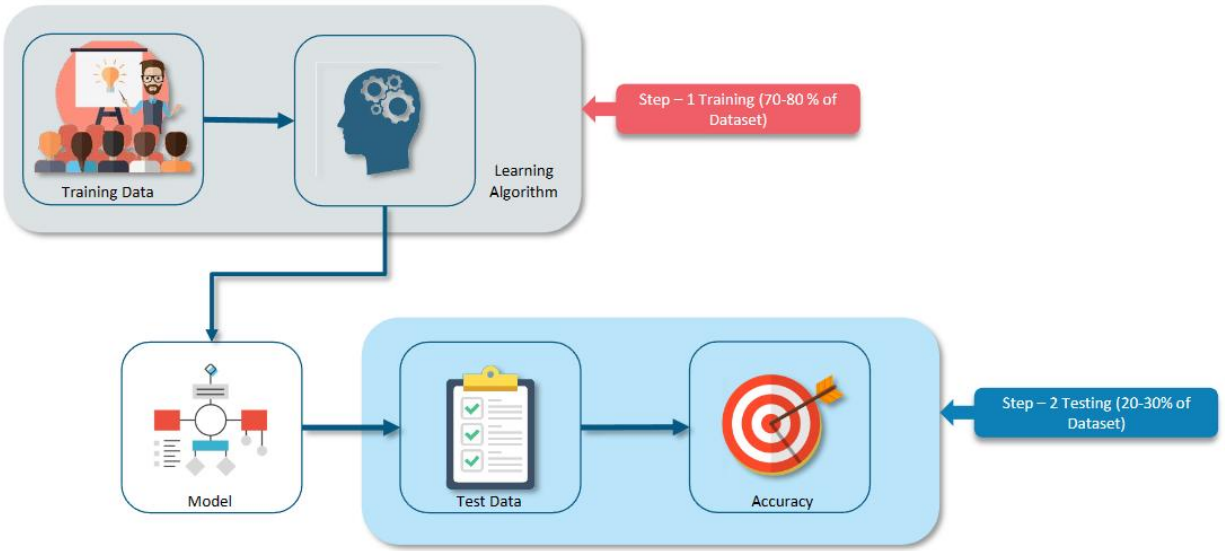
Analyzing data to identify patterns and trends is key to the transportation industry, which relies on making routes more efficient and predicting potential problems to increase profitability

edureka!

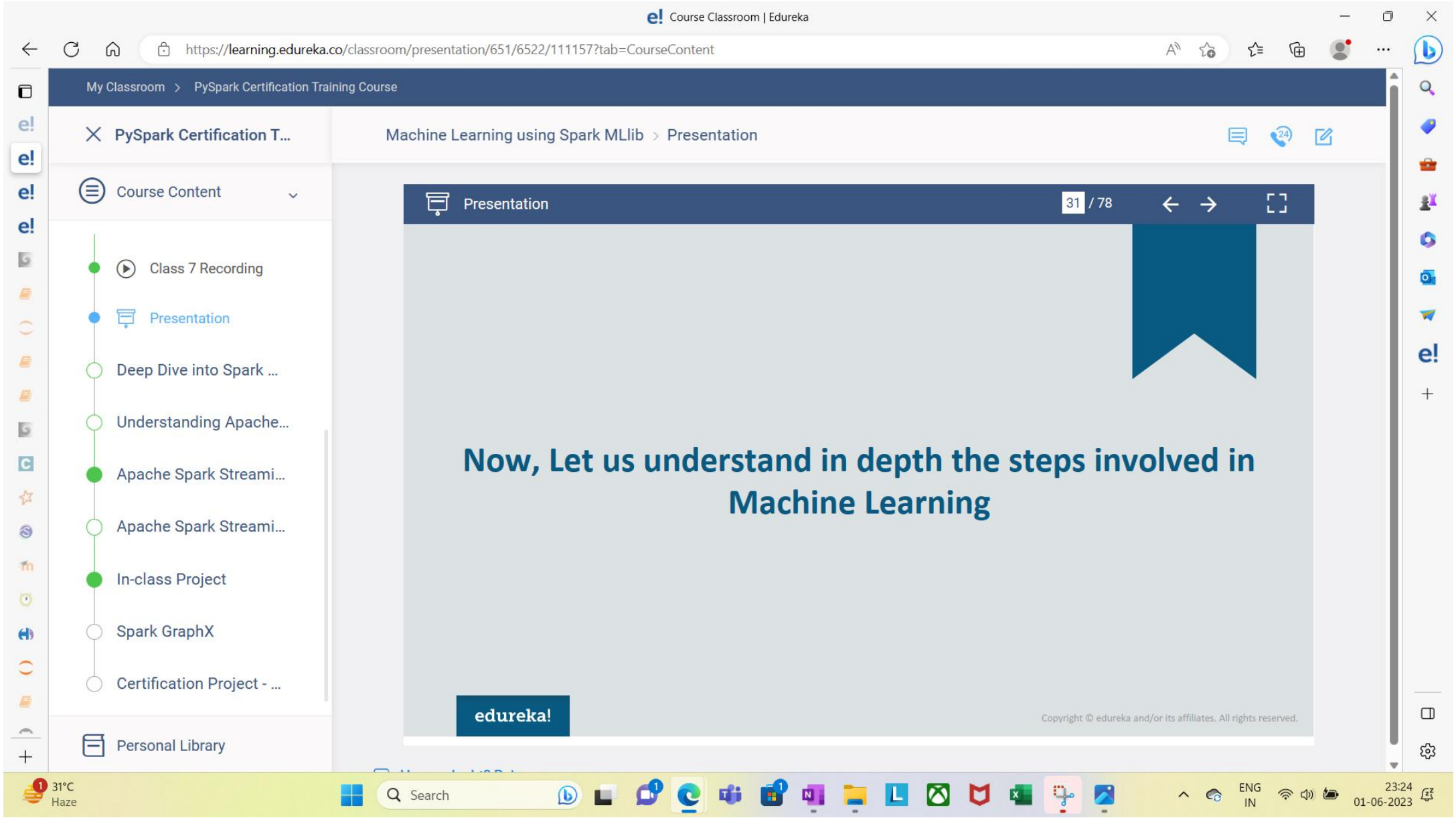

e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
29 / 78
Let's have a look at the Phases of Machine Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 30 / 78

Phases of Typical Machine Learning



edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
31 / 78
Now, Let us understand in depth the steps involved in Machine Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 32 / 78

Steps of Machine Learning

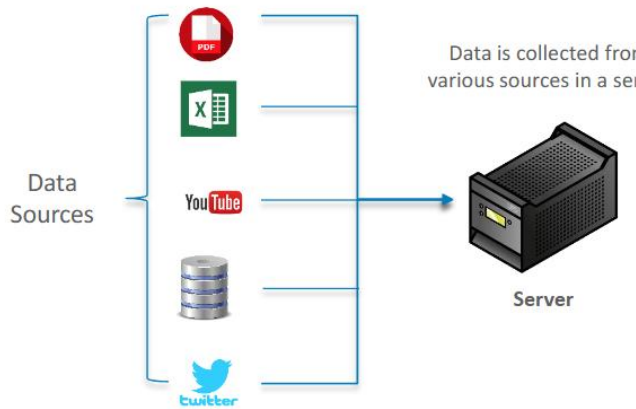
edureka!

Presentation 33 / 78

Collecting Data

This stage involves the collection of all relevant data from various sources

1. Collecting Data
2. Data Wrangling
3. Analyze Data
4. Train Algorithm
5. Test Algorithm
6. Deployment



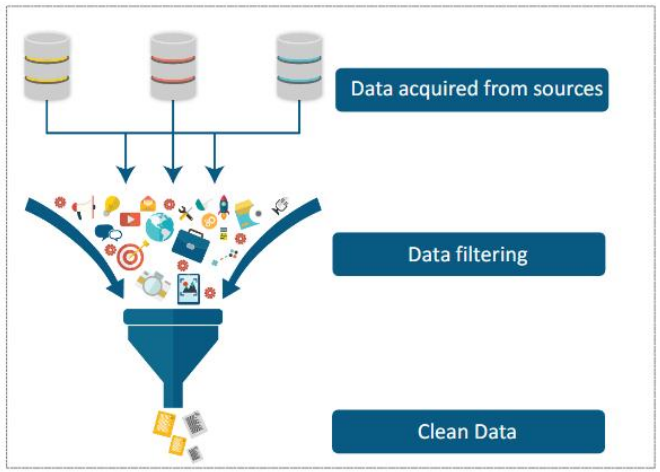
edureka!

Presentation 34 / 78

Data Wrangling

Data Wrangling is the process of cleaning and converting "Raw Data" into a format that allows convenient consumption

1. Collecting Data
2. Data Wrangling
3. Analyze Data
4. Train Algorithm
5. Test Algorithm
6. Deployment



edureka!

Presentation 35 / 78

Analyze Data

1. Collecting Data
2. Data Wrangling
3. Analyze Data
4. Train Algorithm
5. Test Algorithm
6. Deployment

Data is analysed to select and filter the data required to prepare the model

Model

Feature selection

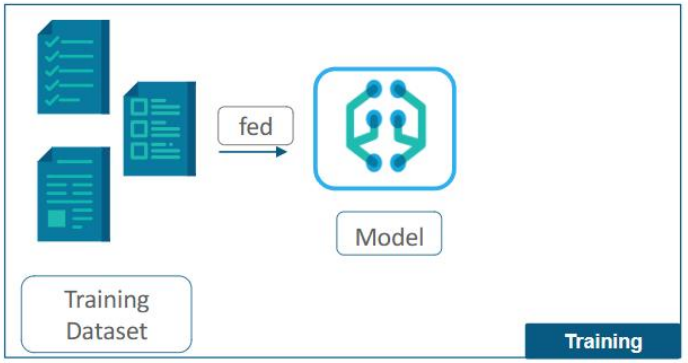
edureka!

Presentation 36 / 78

Train Algorithm

1. Collecting Data
2. Data Wrangling
3. Analyze Data
4. Train Algorithm
5. Test Algorithm
6. Deployment

The algorithm is trained on the training dataset, through which algorithm understands the pattern and the rules which govern the data



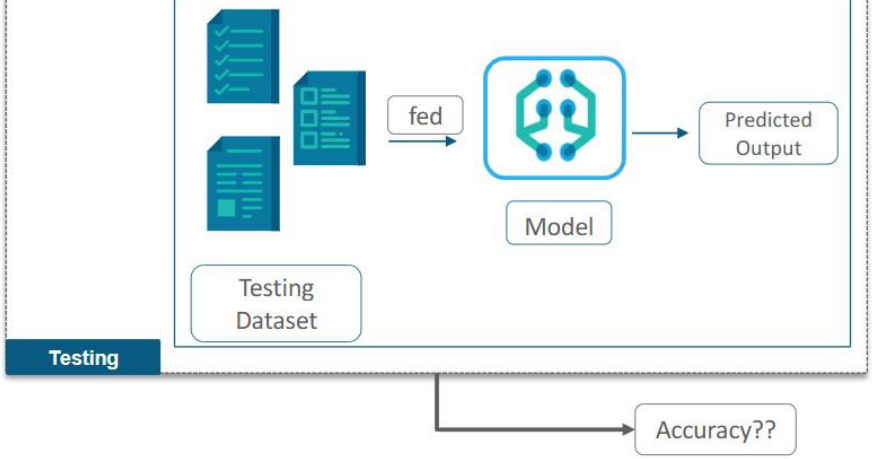
edureka!

Presentation 37 / 78

Test Algorithm

1. Collecting Data
2. Data Wrangling
3. Analyze Data
4. Train Algorithm
5. Test Algorithm
6. Deployment

The testing dataset determines the accuracy of our model



edureka!

Presentation 38 / 78

Deployment

1. Collecting Data
2. Data Wrangling
3. Analyze Data
4. Train Algorithm
5. Test Algorithm
6. Deployment

If the speed and accuracy of the model is acceptable, then that model should be deployed in the real system

The model that is used in production should be made with all the available data

Models improve with the amount of available data used to create the model

The results of the model need to be incorporated in the business strategy

edureka!

Presentation 39 / 78

Deployment

1. Collecting Data
2. Data Wrangling
3. Analyze Data
4. Train Algorithm
5. Test Algorithm
6. Deployment

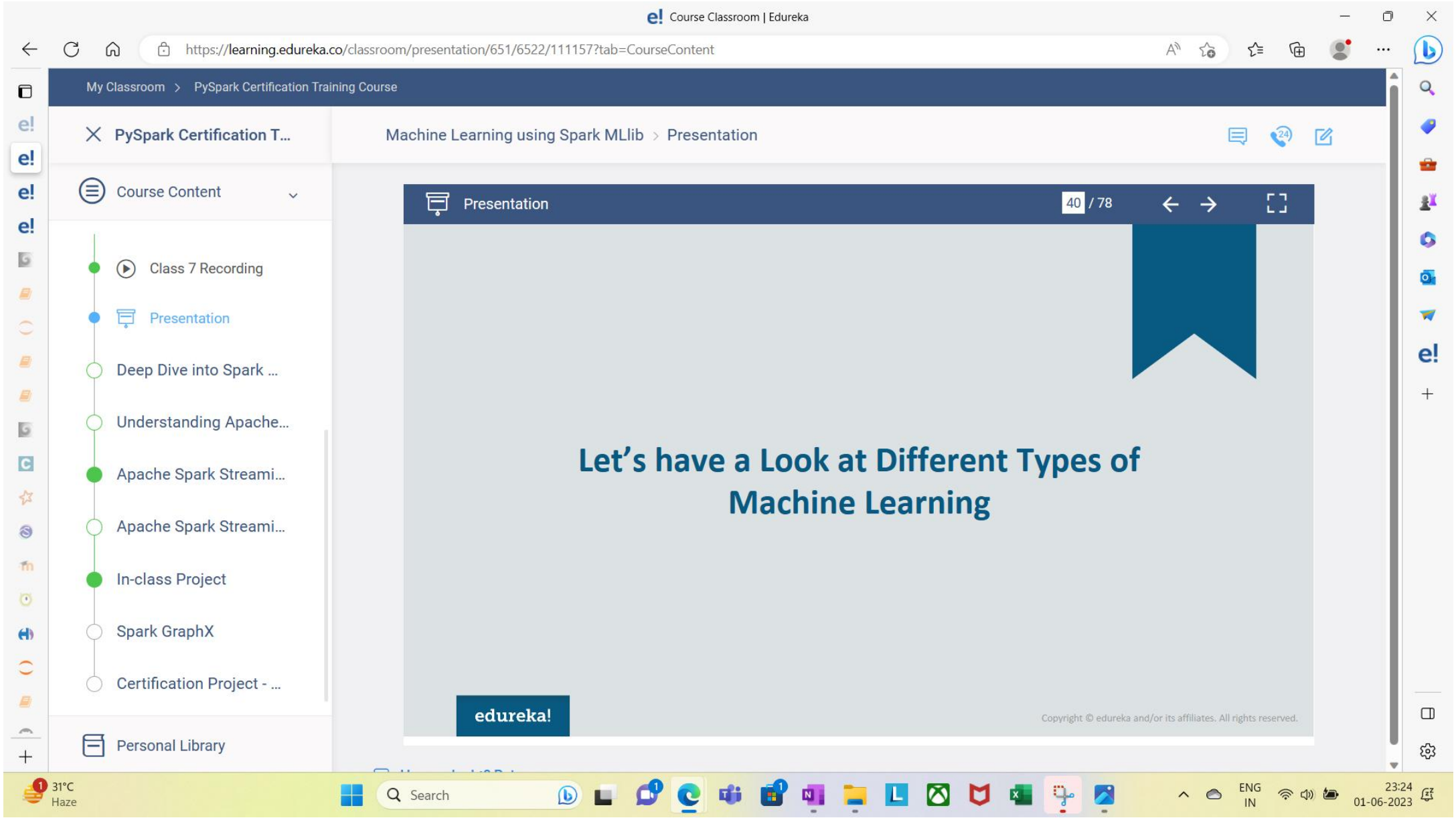
After the model is deployed based upon it's performance the model is updated and improved, if there is a dip in performance the model is retrained

The model that is used in production should be made with all the available data

Models improve with the amount of available data used to create the model

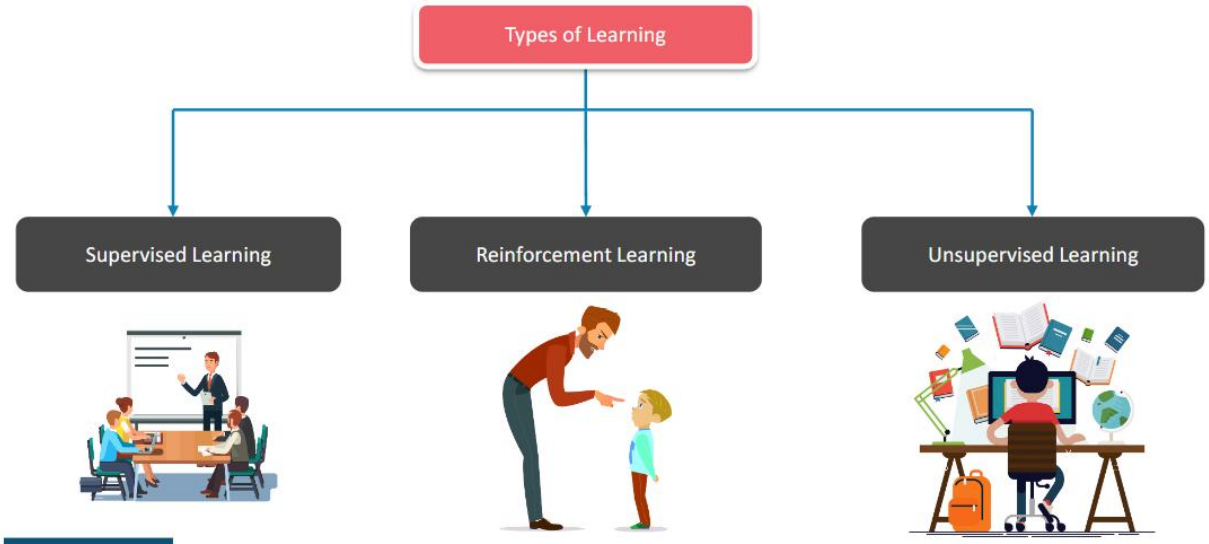
The results of the model need to be incorporated in the business strategy

edureka!

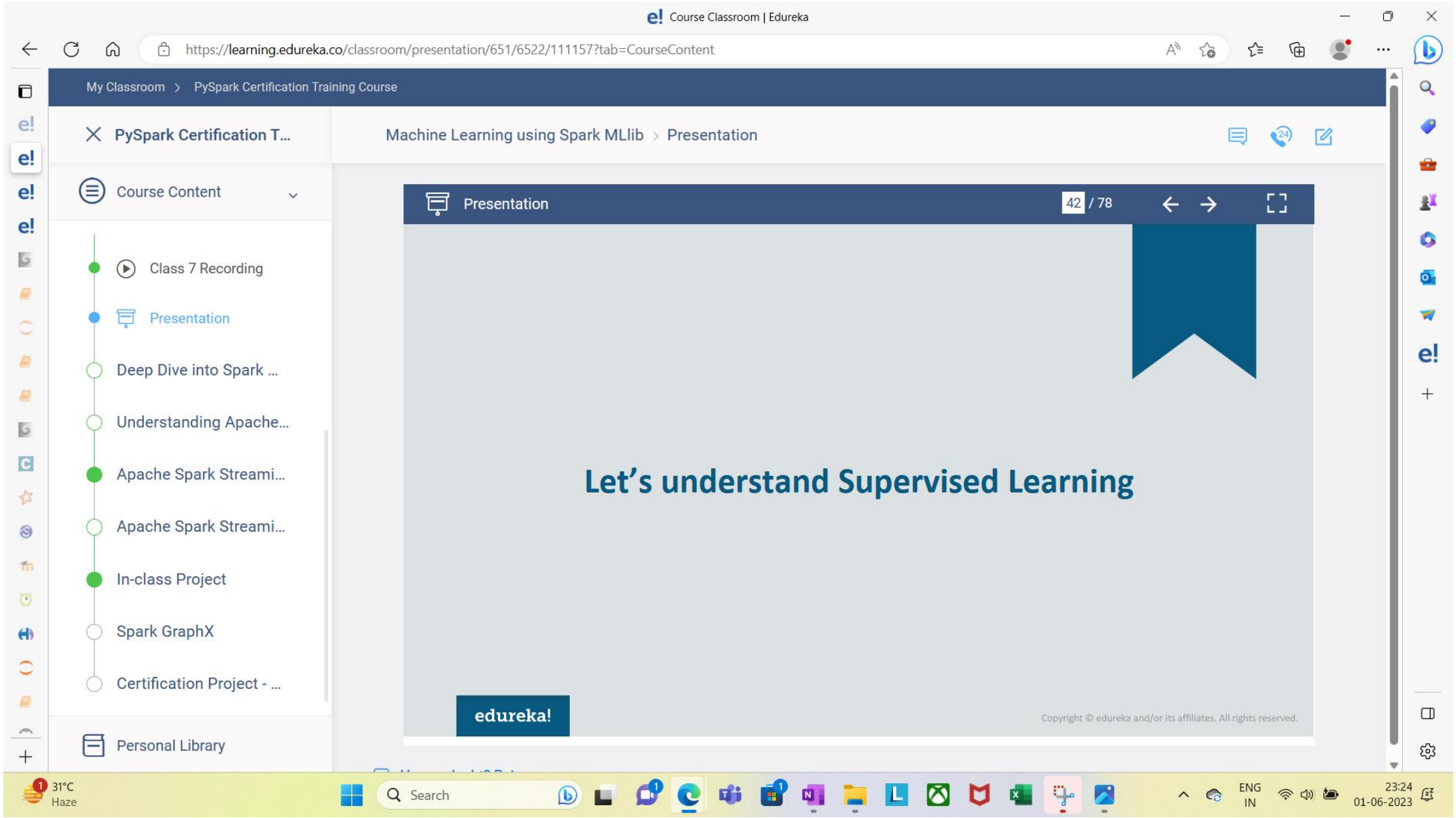

e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
40 / 78
Let's have a Look at Different Types of Machine Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 41 / 78

Types of Machine Learning

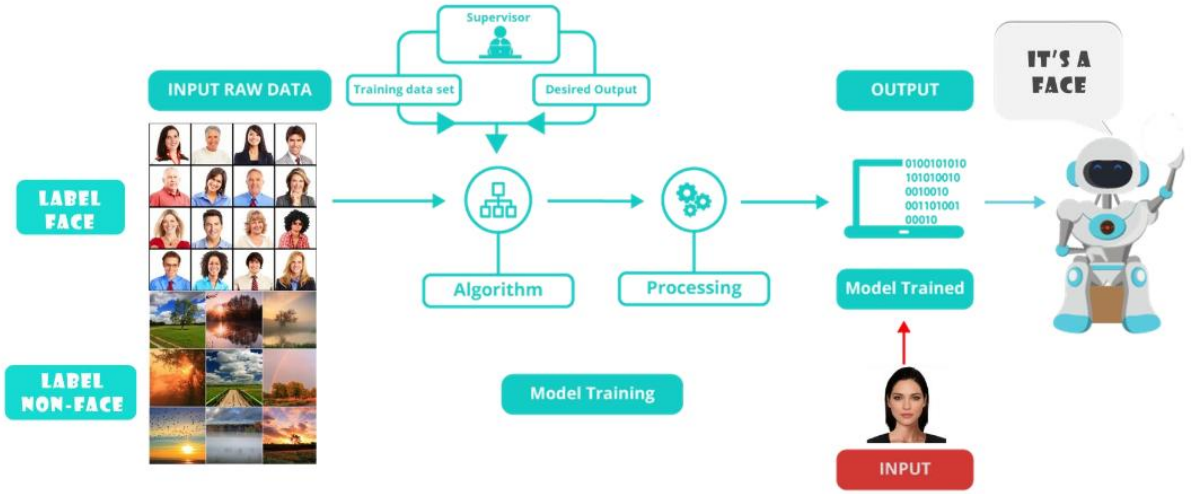


edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
42 / 78
Let's understand Supervised Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 43 / 78

Supervised Learning



edureka!

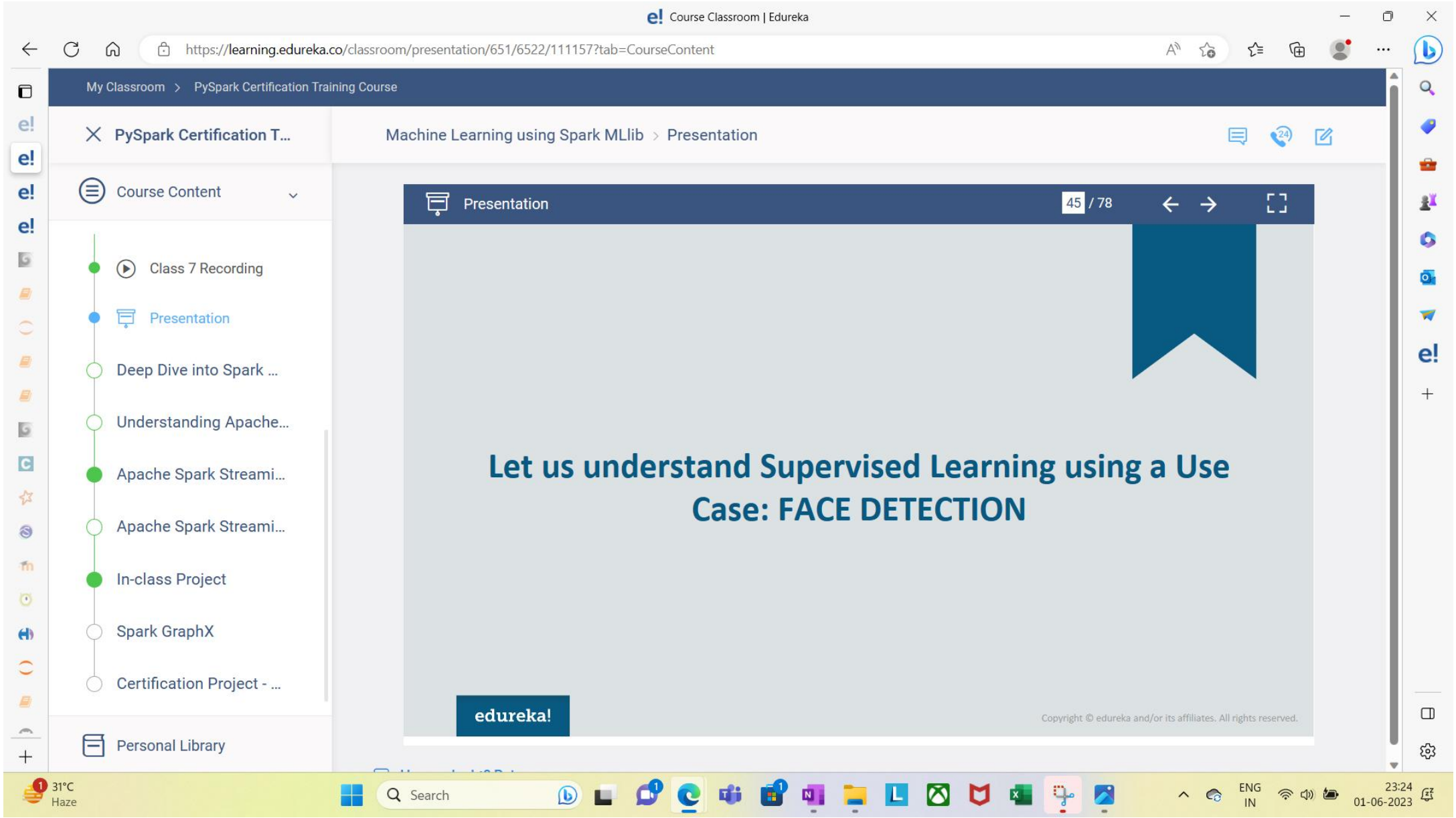
Presentation 44 / 78

Supervised Learning

- **Supervised learning** algorithms are *trained using labelled* examples, such as an *input where the desired output is known*
- The *learning algorithm receives* a set of *inputs* along with the *corresponding* correct *outputs*, and the *algorithm learns* by *comparing* its *actual output with correct outputs* to find errors
- It then *modifies* the *model accordingly*, through methods like *classification*, *regression*, *prediction* and *gradient boosting*
- *Supervised learning* uses patterns to predict the values of the *label on additional unlabelled data*

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.

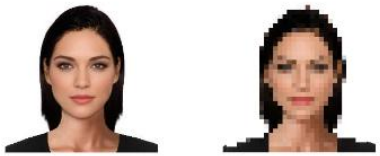
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
45 / 78
Let us understand Supervised Learning using a Use Case: FACE DETECTION
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

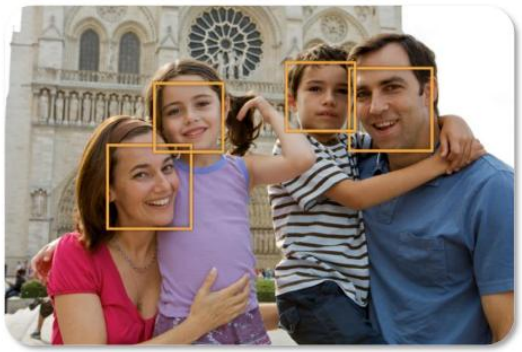
Presentation 46 / 78

Face Detection

- What is a picture? But a collection of pixels. Right ?



- The Question arises "***How can we detect faces***?"
- The human brain has been taught what a face looks like, so it can easily distinguish faces from non faces
- So we use the same approach here, i.e. **TEACH THE MACHINE**

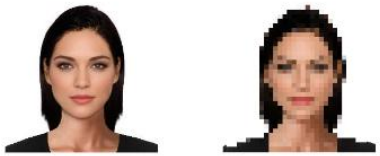


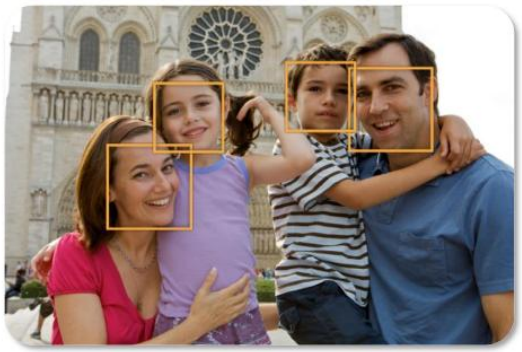
edureka!

Presentation 47 / 78

Face Detection

- What is a picture? But a collection of pixels. Right ?





- The Question arises "***How can we detect faces***?"
- The human brain has been taught what a face looks like, so it can easily distinguish faces from non faces
- So we use the same approach here, i.e. **TEACH THE MACHINE**

This is what we do in *Supervised Learning*

edureka!

Presentation 48 / 78

Providing Input to Machine With Labels

We have a *collection of photos (Faces and Non-Faces)*, out of that we take ***70-80%*** of images and provide as input with Labels to the machine

Label : **"Face"**

Label : **"Non-Face"**

Input to Machine

The *machine scans* the *images* and *finds all* the *pixel features* of a picture that are particular to faces and *creates* a ***MODEL***.

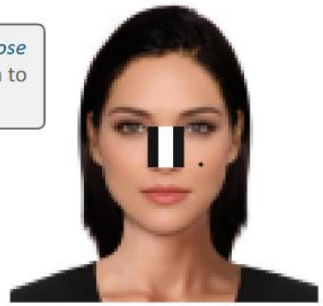
edureka!

Presentation 49 / 78

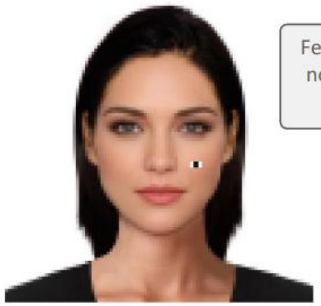
Relevant and Irrelevant Features

- We have more than ***160,000*** features within a window to be calculated
- We don't need all the features, rather just the useful ones
- So using *Machine Learning* it makes a *collection of strong features* and are finally *combined* to *form* a *MODEL*

Features like *Eyes, nose and lips* are common to a face



Relevant Feature



Irrelevant Feature

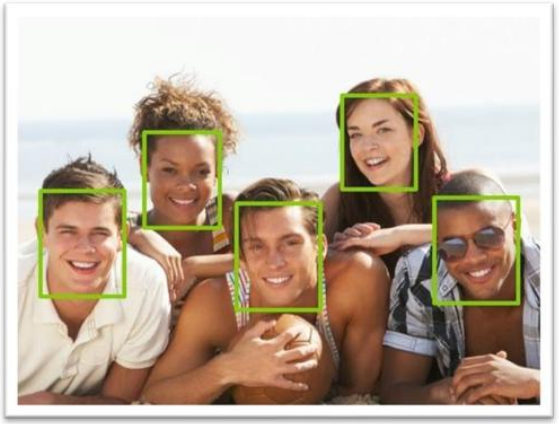
Features like *moles* are not common features of a face

edureka!

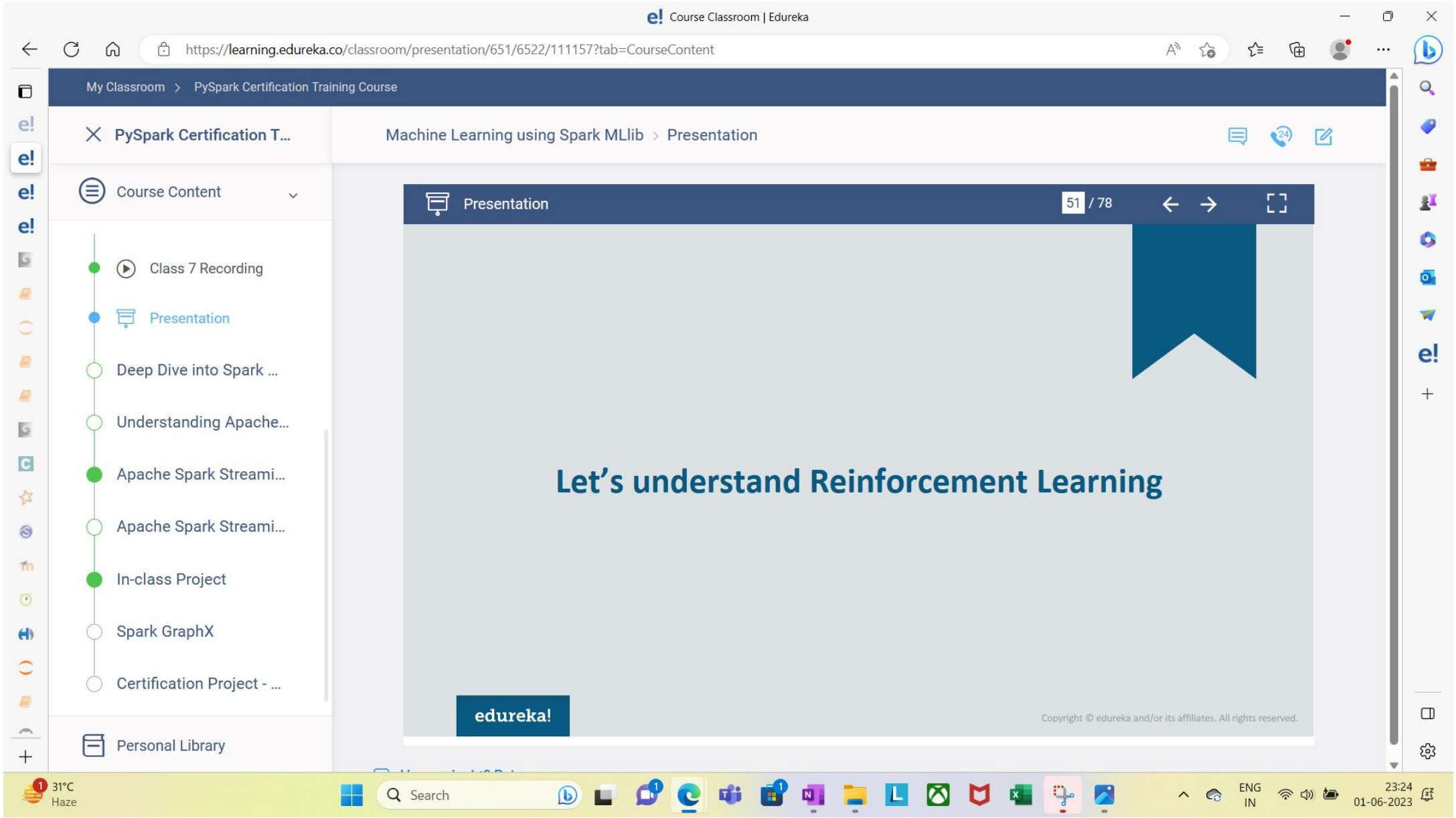
Presentation 50 / 78

Getting Results

- After the model is created, the *remaining 20-30%* of the *images* are *used for testing* purposes
- These *images* are *provided to* the *model* to *get* the *desired results*

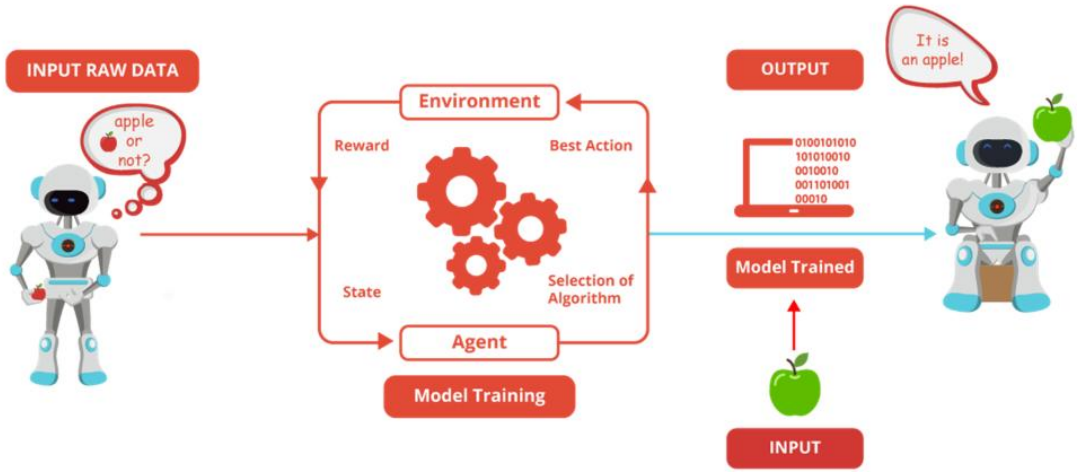


edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
51 / 78
Let's understand Reinforcement Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 52 / 78

Reinforcement Learning

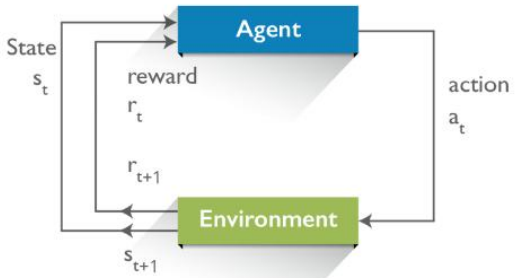


edureka!

Reinforcement Learning

- **Reinforcement** learning is *learning how to map situations to actions* so as to *maximize a reward* and often used for robotics, gaming and navigation
- With reinforcement learning, the *algorithm discovers through trial and error* which actions yield the greatest rewards
- The algorithm provides information about whether the answer is correct or not, but does not tell how to improve it

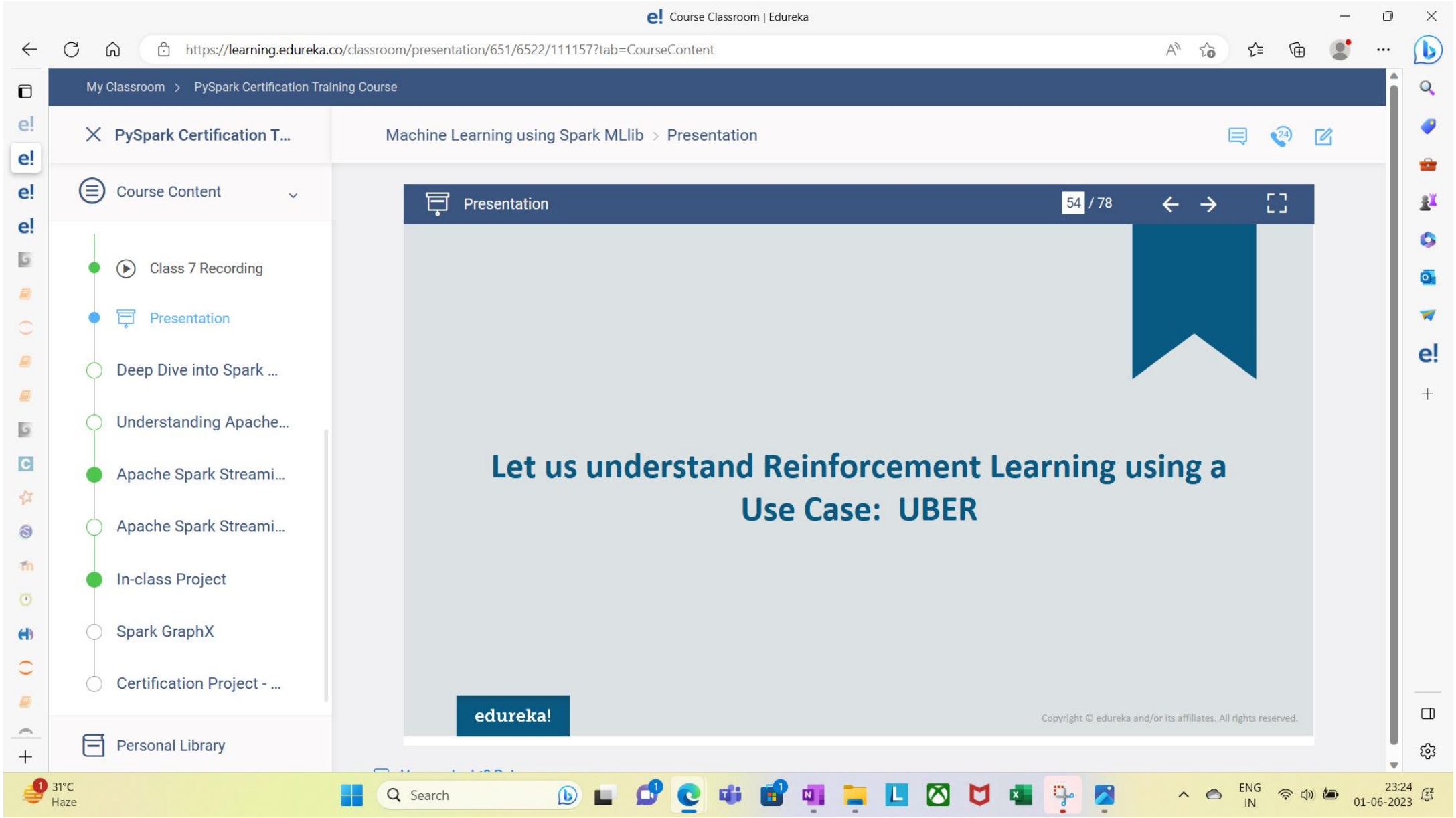
actions are what the agent can do



The **environment** is everything the agent interacts with

The **agent** is the *learner* or *decision maker whose job* is to *choose actions that maximize the expected reward* over a given amount of time

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
54 / 78
Let us understand Reinforcement Learning using a Use Case: UBER
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 55 / 78

Self Driving Cars at UBER : USE CASE

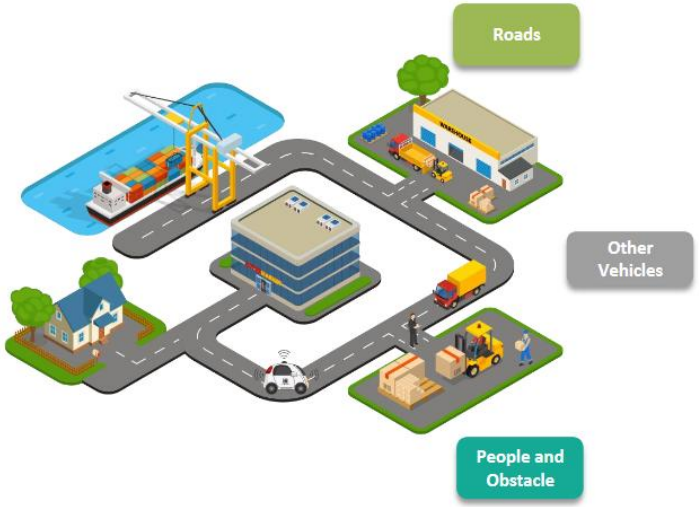
AGENT

edureka!

Presentation 56 / 78

Self Driving Cars at UBER : USE CASE

ENVIRONMENT



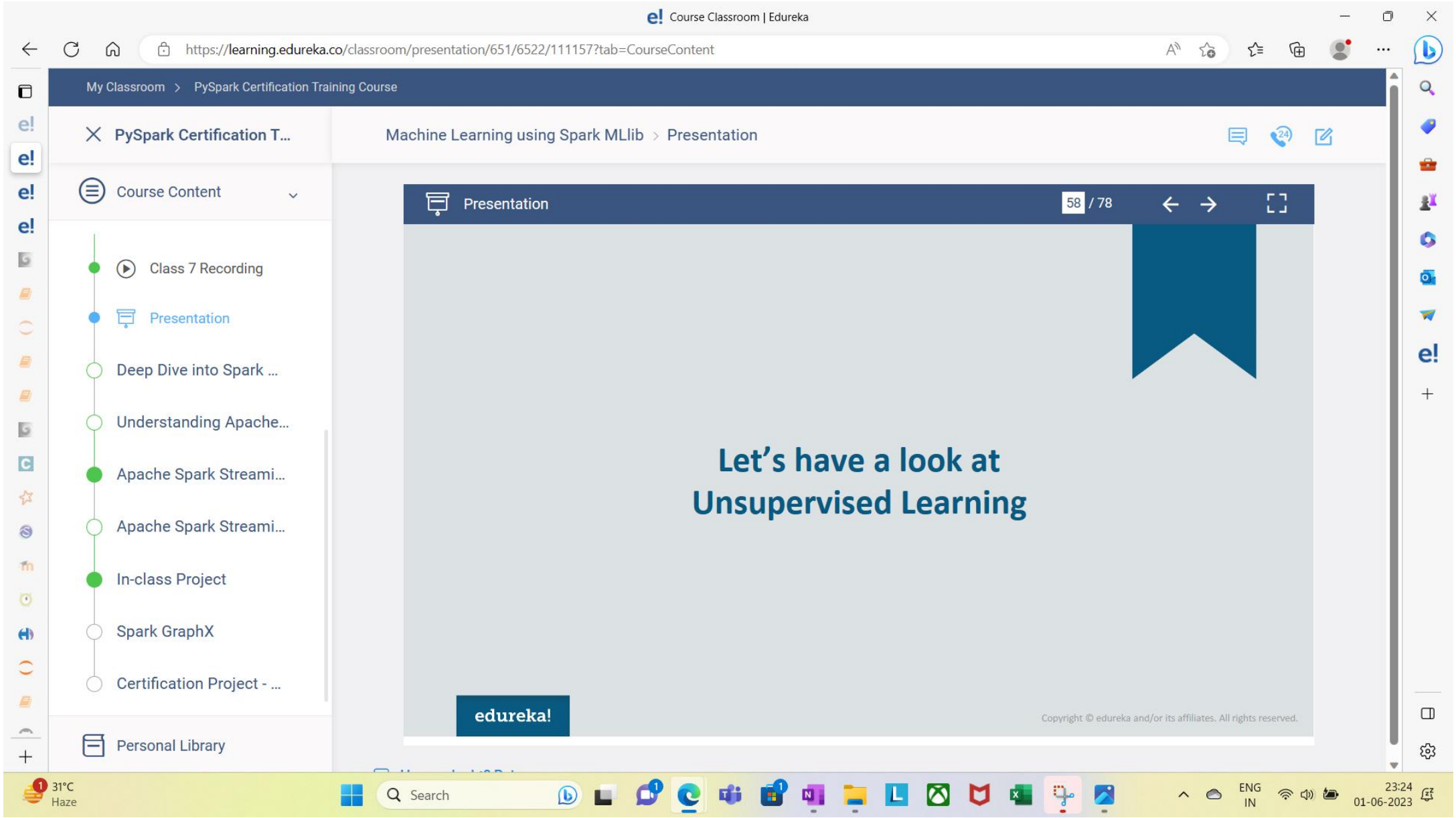
edureka!

Presentation 57 / 78

Self Driving Cars at UBER : USE CASE

ACTIONS and REWARD

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
58 / 78
Let's have a look at
Unsupervised Learning
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

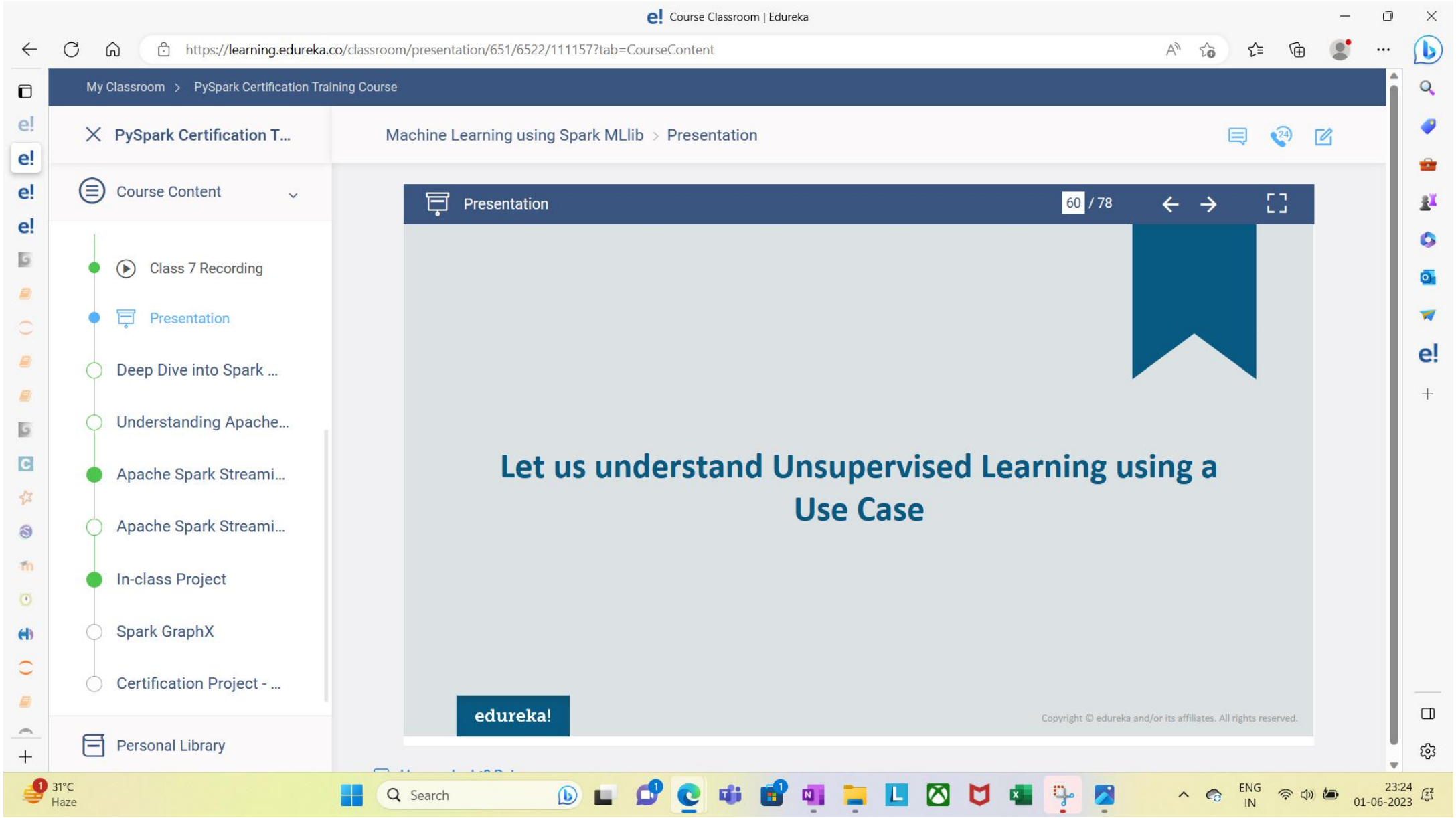
Presentation 59 / 78

Unsupervised Learning

- **Unsupervised Learning** is used against data that has no historical labels. The system is not told the "right answer." **The algorithm must figure out what is being shown**
- The goal is to explore the data and find some structure within. Unsupervised learning works well on transactional data
- Popular techniques include *self-organizing maps, nearest-neighbour mapping, K-MEANS CLUSTERING and singular value decomposition*
- These algorithms are also used to segment text topics, recommend items and identify data outliers

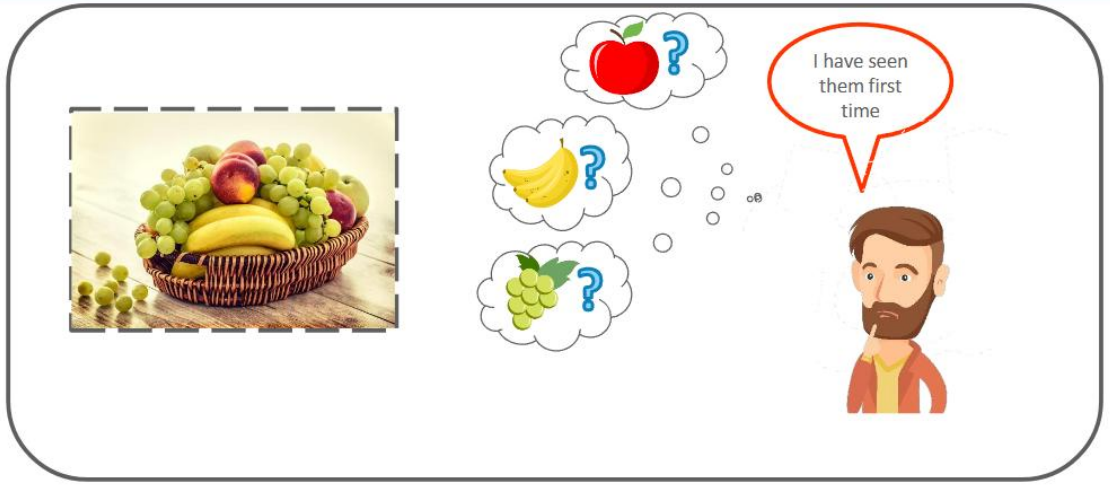
The Machine learns the patterns all by itself and divides them into categories. There are no labels provided in advance

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
60 / 78
Let us understand Unsupervised Learning using a Use Case
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 61 / 78

Unsupervised Learning: USE CASE



edureka!

Presentation 62 / 78

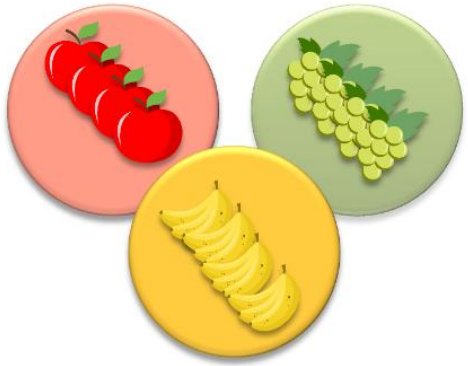
Unsupervised Learning: USE CASE

What will you do first???

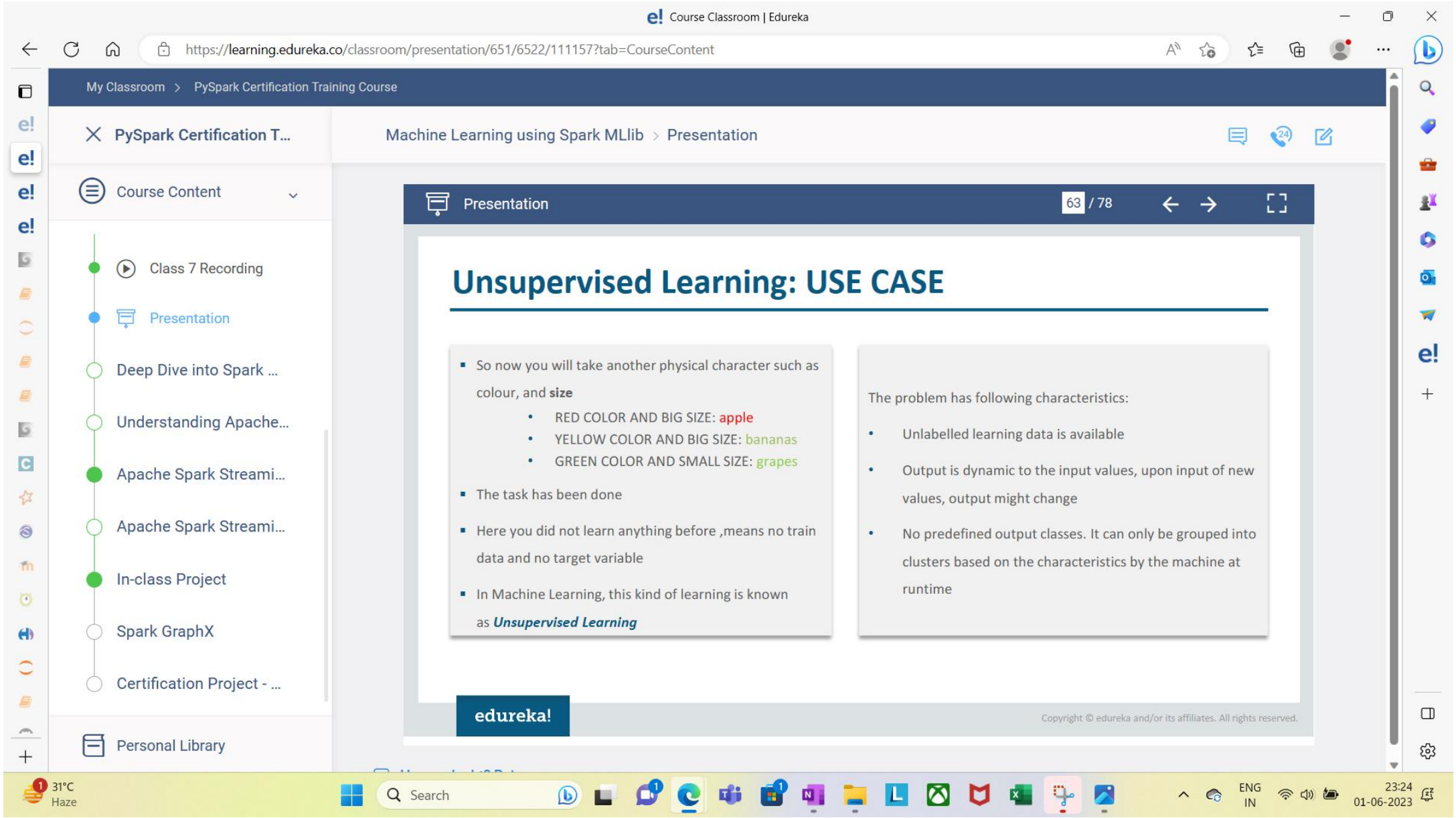
You will take a fruit and you will arrange them by considering the physical character of that particular fruit

Suppose you have considered colour

- Then you will arrange them on considering base condition as **colour**
- Then the groups will be something like this
  - Red colour group: apples
  - Green colour group: grapes
  - Yellow colour group: bananas

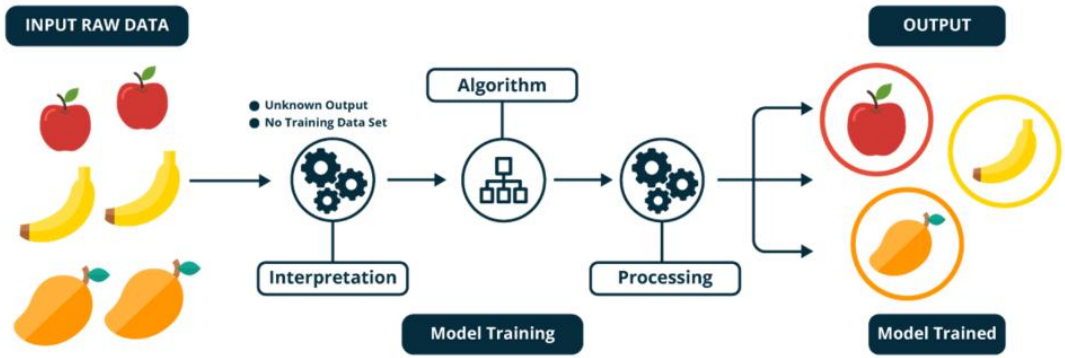


edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
63 / 78
Unsupervised Learning: USE CASE
So now you will take another physical character such as colour, and size
RED COLOR AND BIG SIZE: apple
YELLOW COLOR AND BIG SIZE: bananas
GREEN COLOR AND SMALL SIZE: grapes
The task has been done
Here you did not learn anything before ,means no train data and no target variable
In Machine Learning, this kind of learning is known as Unsupervised Learning
The problem has following characteristics:
Unlabelled learning data is available
Output is dynamic to the input values, upon input of new values, output might change
No predefined output classes. It can only be grouped into clusters based on the characteristics by the machine at runtime
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C Haze
Search
ENG IN
23:24 01-06-2023

Presentation 64 / 78

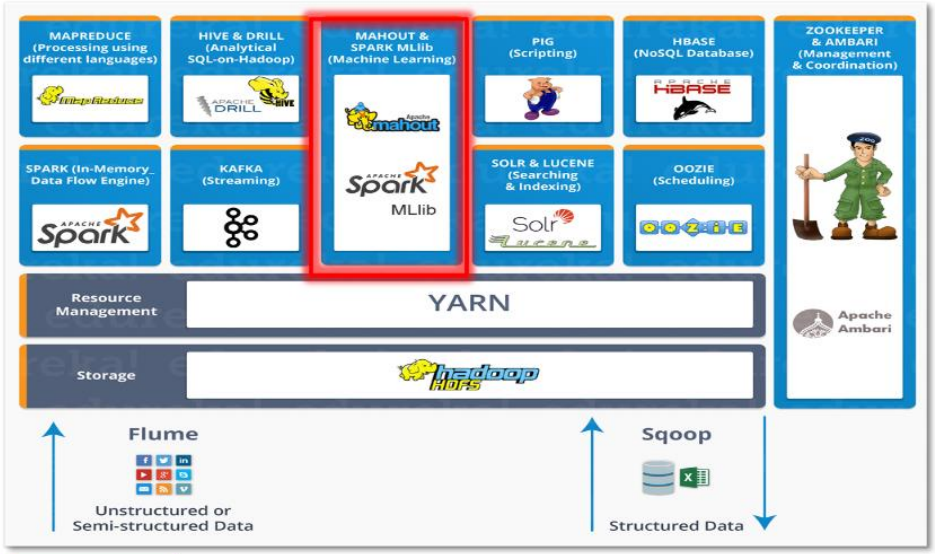
Unsupervised Learning: USE CASE



edureka!

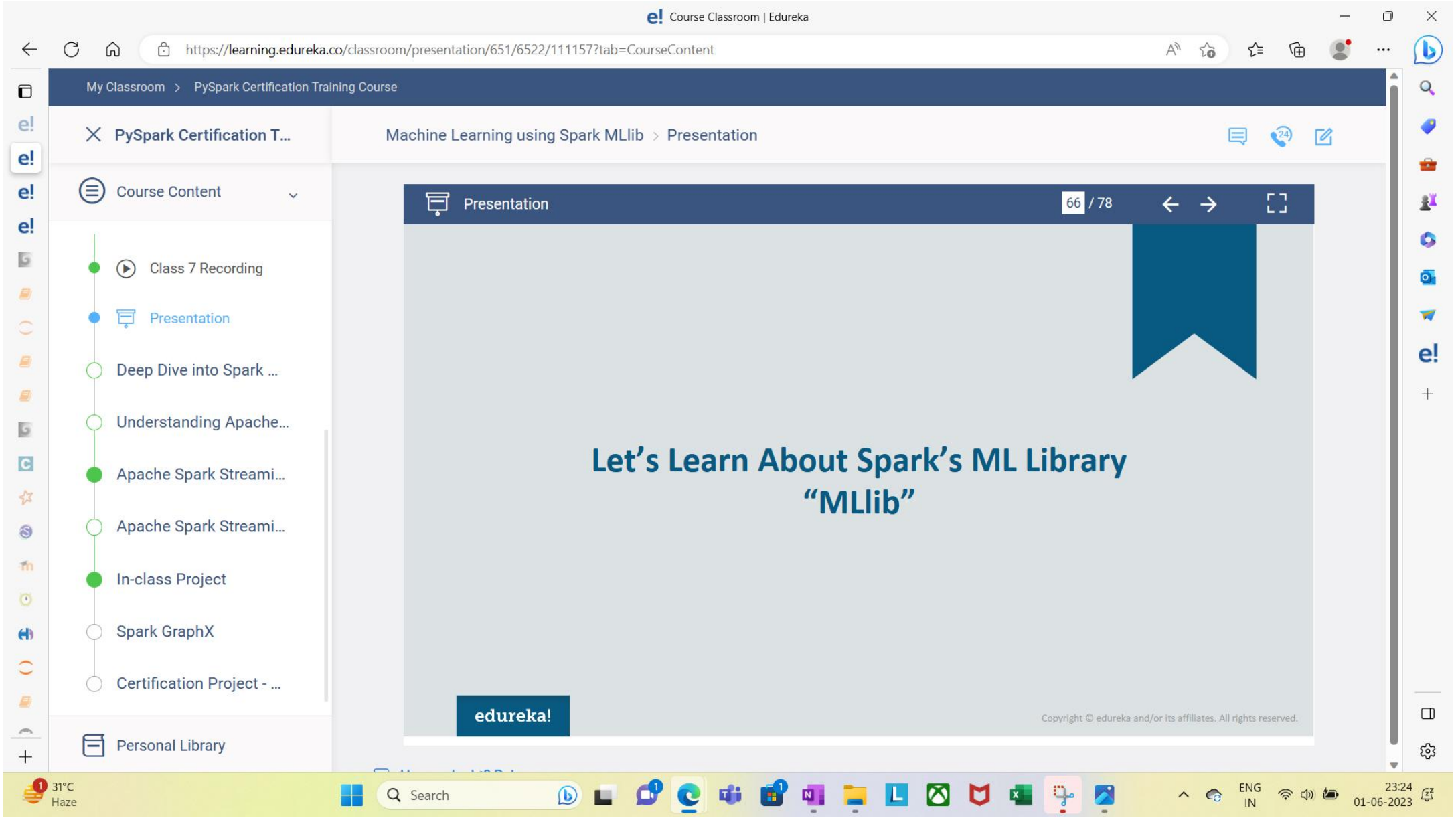
Presentation 65 / 78

Recap of the Topics



Let us recall the concepts of Machine Learning that we have learnt, since they will be used in further module where we will be working on Spark MLlib

edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
66 / 78
Let's Learn About Spark's ML Library
"MLlib"
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 67 / 78

MLlib

- MLlib is *Spark's machine learning (ML) library*
- It consists of *common learning algorithms* and *utilities*, including *classification*, *regression*, *clustering*, *collaborative filtering*, *dimensionality reduction*, as well as *lower-level optimization primitives* and *higher-level pipeline APIs*
- It divides into two packages:
  - ***spark.mllib*** contains the original API built on top of ***RDDs***
  - ***spark.ml*** provides higher-level API built on top of ***DataFrames*** for constructing ML pipelines

edureka!

Presentation 68 / 78

Why MLlib?

Simplicity and Compatibility: Simple APIs familiar to data scientists coming from tools like R and Python

Scalability: Ability to run the same ML code on your laptop and on a big cluster seamlessly without breaking down

Streamlined end-to-end: Building MLlib on top of Spark makes it possible to tackle distinct needs with a single tool instead of many disjointed ones

Faster Processing: Run Programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk

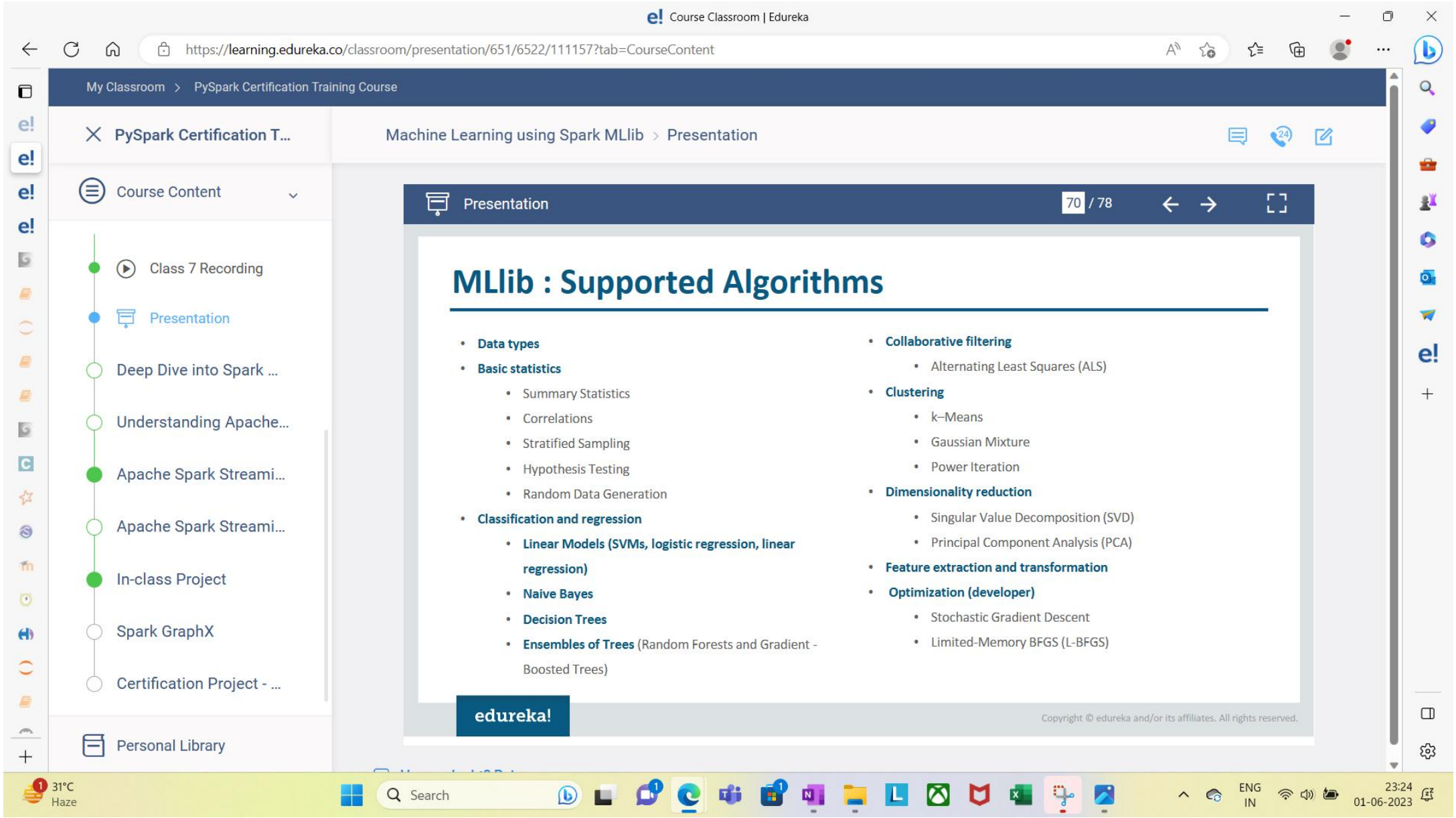
edureka!

Presentation 69 / 78

Spark MLlib Tools

- ***ML Algorithms*:** ML Algorithms form the core of MLlib. These include common learning algorithms such as classification, regression, clustering and collaborative filtering
- ***Featurization*:** Featurization includes feature extraction, transformation, dimensionality reduction and selection
- ***Pipelines*:** Pipelines provide tools for constructing, evaluating and tuning ML Pipelines
- ***Persistence*:** Persistence helps in saving and loading algorithms, models and Pipelines
- ***Utilities*:** Utilities for linear algebra, statistics and data handling

edureka!


e! Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
70 / 78
MLlib : Supported Algorithms
Data types
Basic statistics
Summary Statistics
Correlations
Stratified Sampling
Hypothesis Testing
Random Data Generation
Classification and regression
Linear Models (SVMs, logistic regression, linear regression)
Naive Bayes
Decision Trees
Ensembles of Trees (Random Forests and Gradient - Boosted Trees)
Collaborative filtering
Alternating Least Squares (ALS)
Clustering
k–Means
Gaussian Mixture
Power Iteration
Dimensionality reduction
Singular Value Decomposition (SVD)
Principal Component Analysis (PCA)
Feature extraction and transformation
Optimization (developer)
Stochastic Gradient Descent
Limited-Memory BFGS (L-BFGS)
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C Haze
Search
ENG IN
23:24 01-06-2023

Presentation 71 / 78

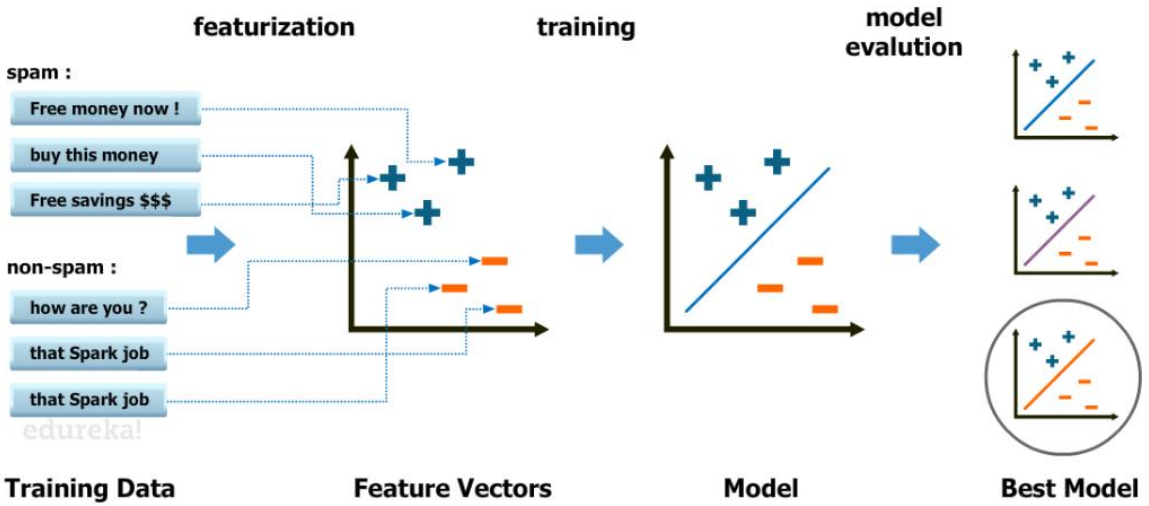
Spark ML Concepts

Concept	Explanation
DataFrame	Spark ML uses DataFrame from Spark SQL as an ML dataset, which can hold a variety of data types. E.g., a DataFrame could have different columns storing text, feature vectors, true labels, and predictions
Transformer	A Transformer is an algorithm which can transform one DataFrame into another DataFrame. E.g., an ML model is a Transformer which transforms DataFrame with features into a DataFrame with predictions
Estimator	An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer. E.g., a learning algorithm is an Estimator which trains on a DataFrame and produces a model
Pipeline	A Pipeline chains multiple Transformers and Estimators together to specify an ML workflow
Parameter	All Transformers and Estimators now share a common API for specifying parameters

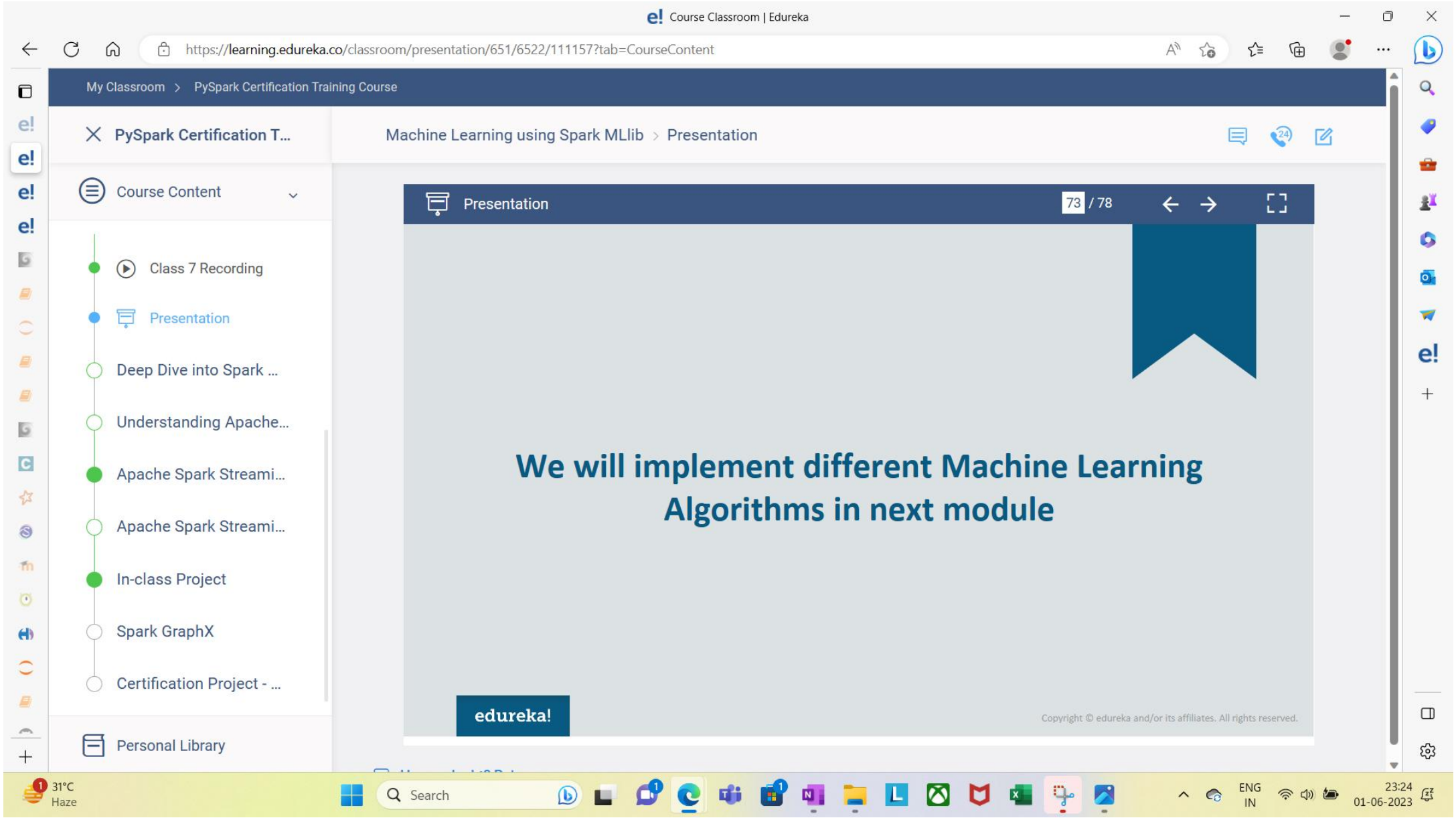
edureka!

Presentation 72 / 78

Typical Steps in ML Pipeline

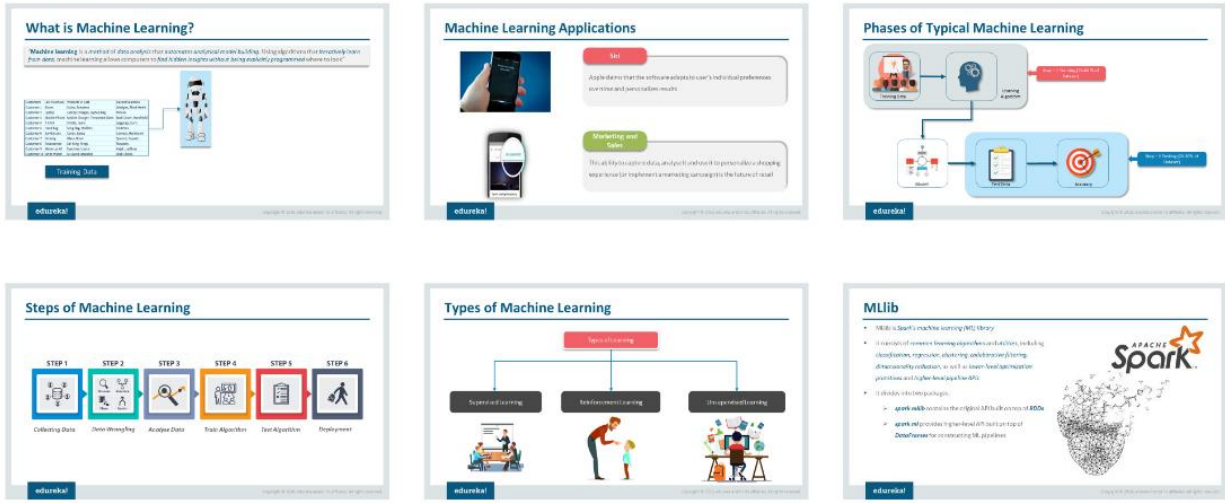


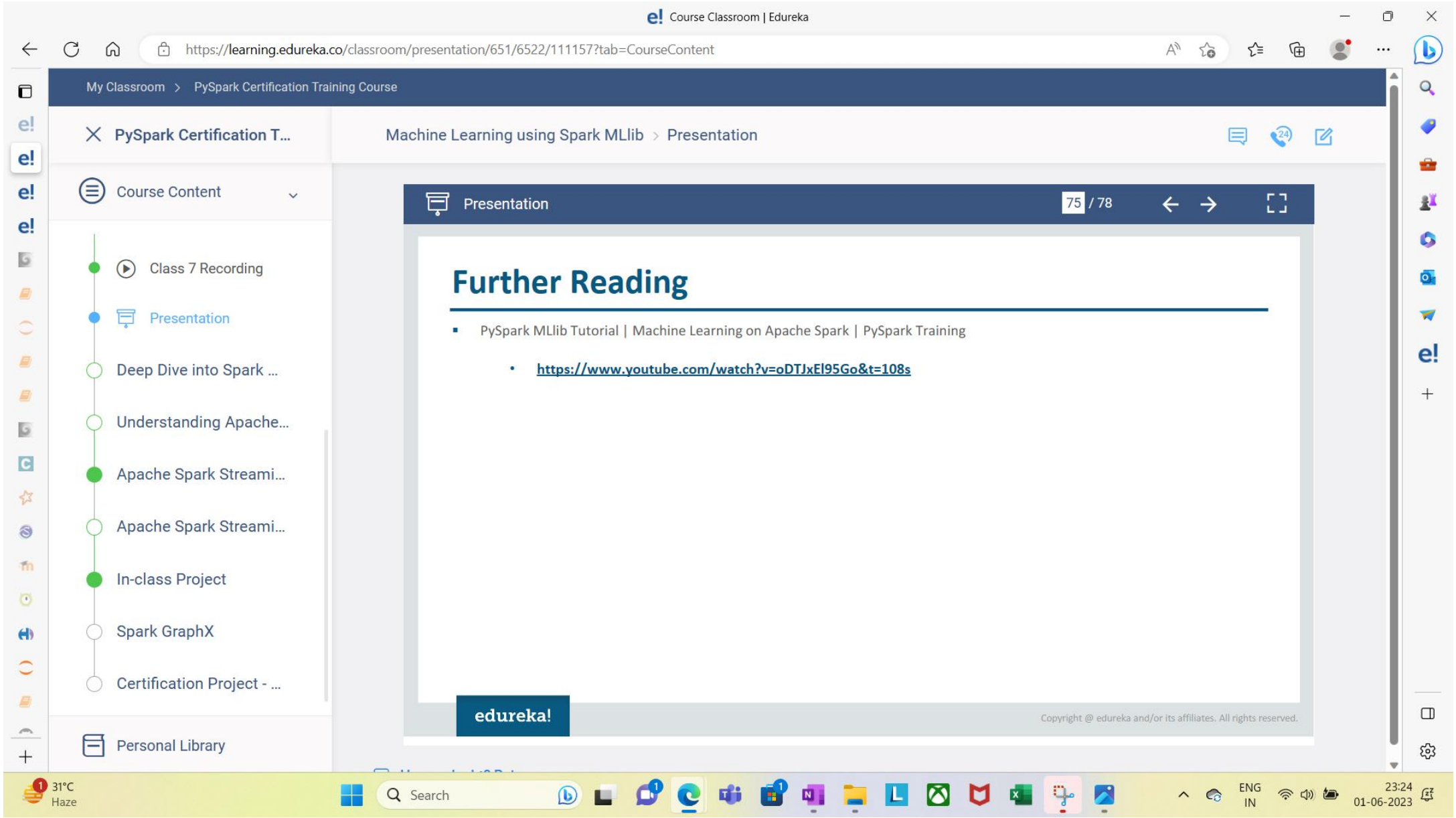
edureka!


Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
73 / 78
We will implement different Machine Learning Algorithms in next module
edureka!
Copyright © edureka and/or its affiliates. All rights reserved.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023

Presentation 74 / 78

Summary




Course Classroom | Edureka
https://learning.edureka.co/classroom/presentation/651/6522/111157?tab=CourseContent
My Classroom > PySpark Certification Training Course
PySpark Certification T...
Machine Learning using Spark MLlib > Presentation
Course Content
Class 7 Recording
Presentation
Deep Dive into Spark ...
Understanding Apache...
Apache Spark Streami...
Apache Spark Streami...
In-class Project
Spark GraphX
Certification Project - ...
Personal Library
Presentation
75 / 78
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.
31°C
Haze
Search
ENG
IN
23:24
01-06-2023