
MACHINE LEARNING

TOPICS

- Machine Learning
- Data Science Workflow
- Supervised Learning

LEARNING OUTCOMES

At the end of this topic, You should be able to,

- Understand the concept of machine learning and supervised algorithm

MACHINE LEARNING

- Machine learning is a subset of artificial intelligence that focuses on developing algorithms and models which can learn from data to make predictions or decisions without being explicitly programmed.

- Instead of being explicitly programmed, machine learning algorithms learn from patterns and data, enabling them to improve over time as they are exposed to more data.

MACHINE LEARNING

The study and construction of programs that learn from repeatedly seeing data, rather than being explicitly programmed by humans.

Machine learning encompasses a wide range of techniques including supervised learning, unsupervised learning, reinforcement learning, and more, each suited to different types of data and tasks.

Type	Dataset
Supervised Learning	Data points have known outcome
Unsupervised Learning	Data points have unknown outcome

MACHINE LEARNING IN OUR DAILY LIVES

Spam Filtering	Web Search	Postal Mail Routing
Fraud Detection	Movie Recommendations	Vehicle Driver Assistance
Web Advertisements	Social Networks	Speech Recognition

MACHINE LEARNING WORKFLOW

MACHINE LEARNING WORKFLOW

Problem Statement	What problem are you trying to solve?
Data Collection	What data do you need to solve it?
Data Exploration & Preprocessing	How should you clean your data so your model can use it?
Feature Engineering	Selecting and extracting meaningful features.
Modeling	Build a model to solve your problem?
Validation	Did I solve the problem?
Decision Making & Deployment	Communicate to stakeholders or put into production?

TYPES OF MACHINE LEARNING

Supervised	data points have known outcome
Unsupervised	data points have unknown outcome

TYPES OF MACHINE LEARNING

Supervised	data points have known outcome
Unsupervised	data points have unknown outcome

MACHINE LEARNING VOCABULARY

- **Target:** predicted category or value of the data (column to predict)
- **Features:** properties of the data used for prediction (non-target columns)
- **Example:** a single data point within the data (one row)
- **Label:** the target value for a single data point

MACHINE LEARNING VOCABULARY

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

MACHINE LEARNING VOCABULARY

Target

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

MACHINE LEARNING VOCABULARY

Features

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

MACHINE LEARNING VOCABULARY

Example →

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

MACHINE LEARNING VOCABULARY

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

Label

SUPERVISED LEARNING

TYPES OF SUPERVISED LEARNING

Regression — outcome is continuous (numerical)

Classification — outcome is a category

TARGET vs. FEATURES

Target: Column to predict

Features: Properties of the data used for prediction (non-target columns)

Features

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

Target

EXAMPLE: SUPERVISED LEARNING PROBLEM

Goal: Predict if an email is spam or not spam.

Data: Historical emails labeled as spam or not spam.

Target: Spam or not spam

Features: Email text, subject, time sent, etc.

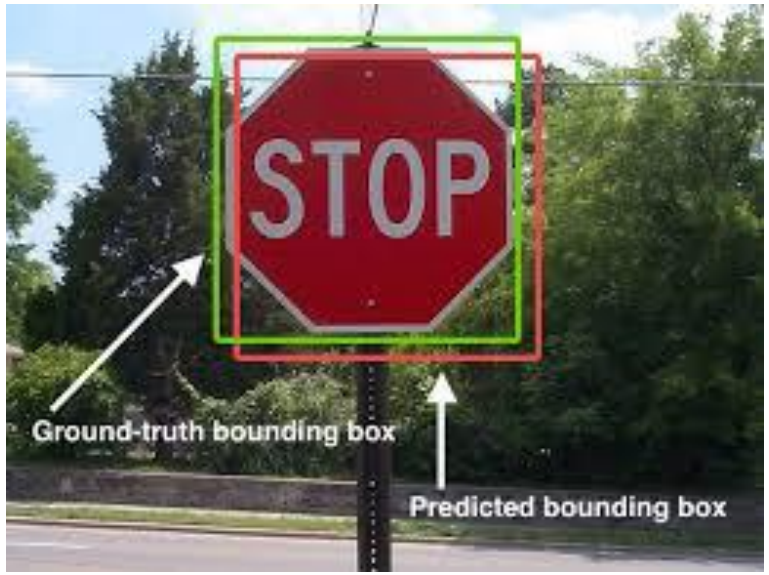
EXAMPLE: SUPERVISED LEARNING PROBLEM

Goal: Predict location of bounding box around an object.

Data: Images with bounding box locations.

Target: Corners of bounding box

Features: Image pixels



FORMULATING A SUPERVISED LEARNING PROBLEM

For a Supervised Learning Problem:

- Collect a labeled dataset (features and target labels).
- Choose the model.
- Choose an evaluation metric:

  "What to use to measure performance."

- Choose an optimization method:[1]

  "How to find the model configuration that gives the best performance."

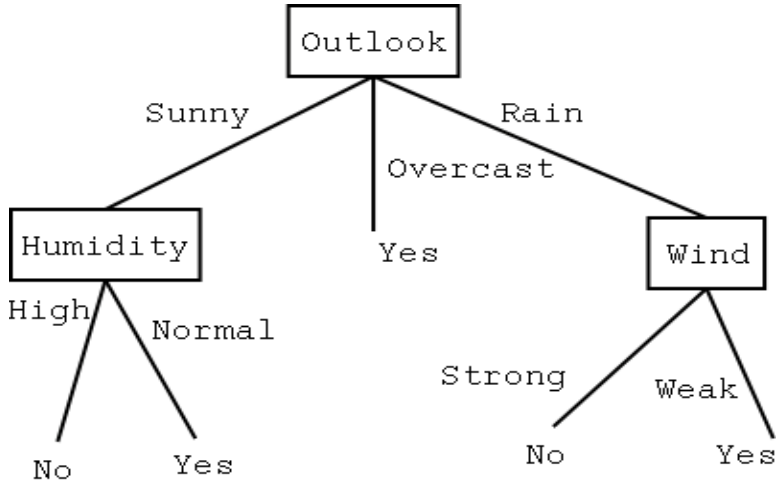
[1] *There are standard methods to use for different models and metrics.*

WHICH MODEL ?

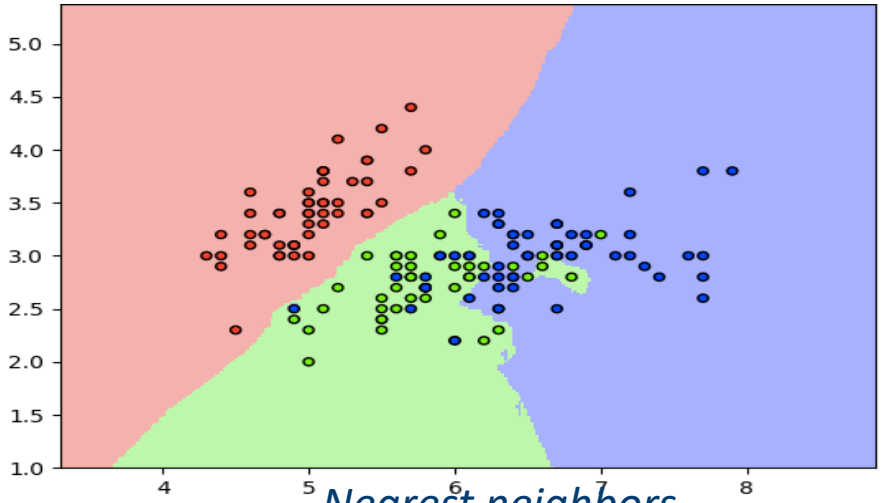
There are many models that represent the problem and make decisions in different ways each with their own advantages and disadvantages.

A **decision tree** makes predictions by asking a series of yes/no questions.

Nearest neighbor makes predictions by having the most similar examples vote.



Decision tree



Nearest neighbors

WHICH MODEL ?

Some considerations when choosing are:

- Time needed for training
- Speed in making predictions
- Amount of data needed
- Type of data
- Problem complexity
- Ability to solve a complex problem
- Tendency to overcomplicate a simple one

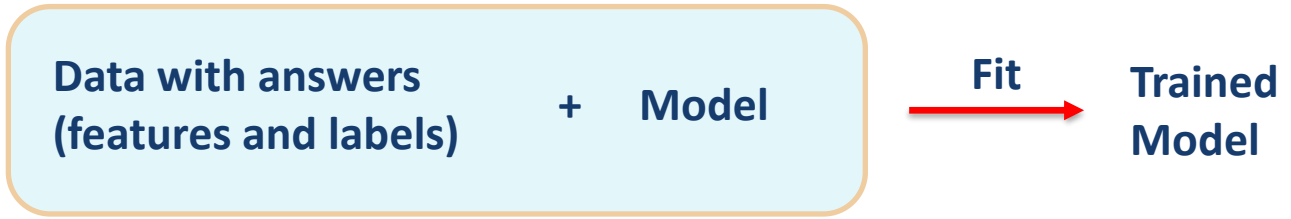
TRAINING

Training Data - The dataset used to train the model.

Optimization - Configures the model for best performance.

SUPERVISED LEARNING OVERVIEW

Training: Train a model with known data.



Inference: Feed unseen data into trained model to make predictions.

SOLUTION: SPLIT DATA INTO TWO SETS

Training Set - Data used during the training process.

Test Set - Data used to measure performance, simulating unseen data[1].

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

Training Set

Testing Set

EVALUATION METRICS IN MACHINE LEARNING

Confusion Matrix
Matrix showing true and false positives/negatives.

Accuracy
Proportion of correctly classified instances.

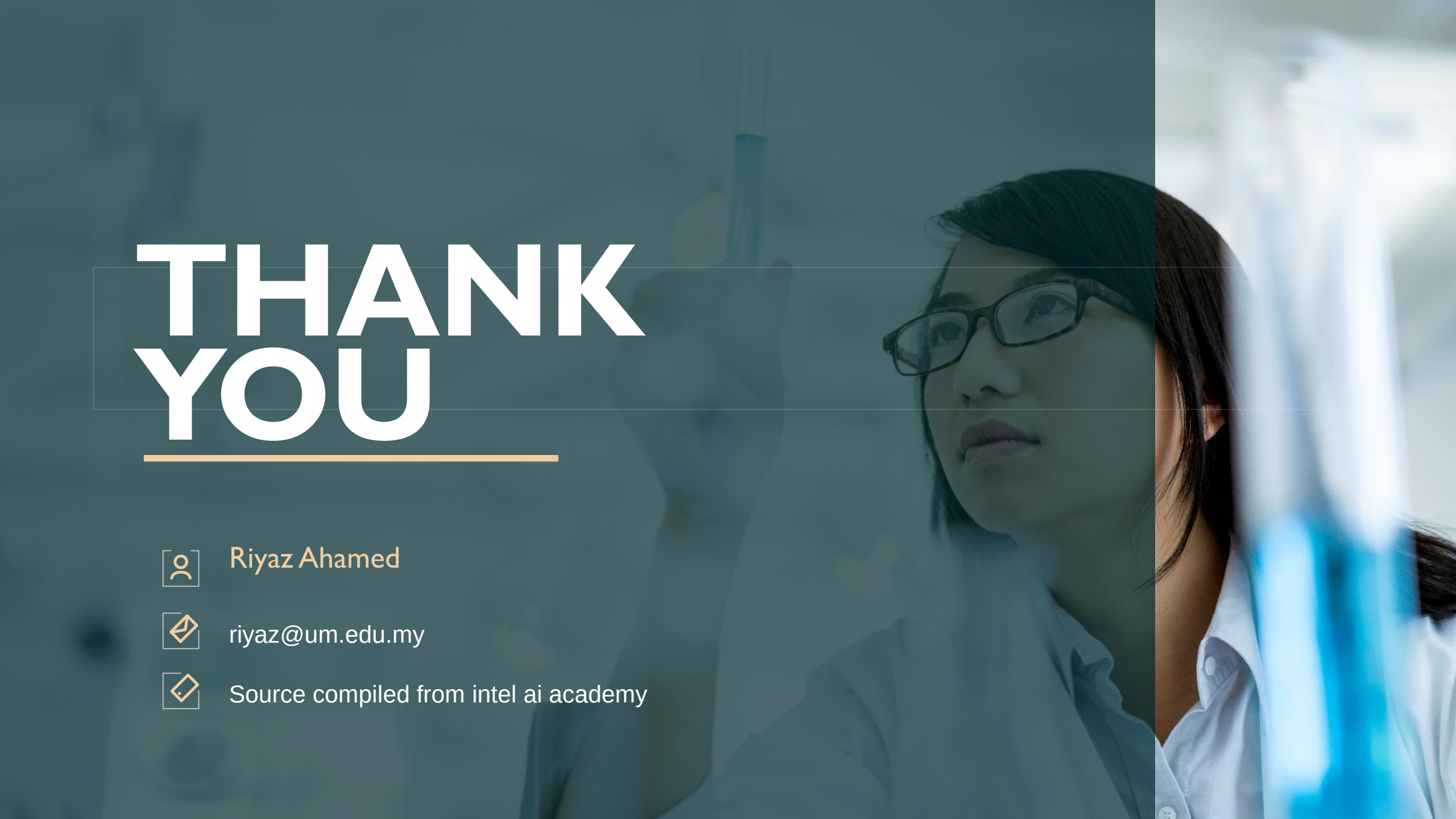
Precision
True positive rate among predicted positives.

Recall
True positive rate among actual positives.

F1 Score
Harmonic mean of precision and recall.

ROC Curve
Plot of true positive rate against false positive rate.

SUMMARY



THANK YOU

Riyaz Ahamed

riyaz@um.edu.my

Source compiled from intel ai academy