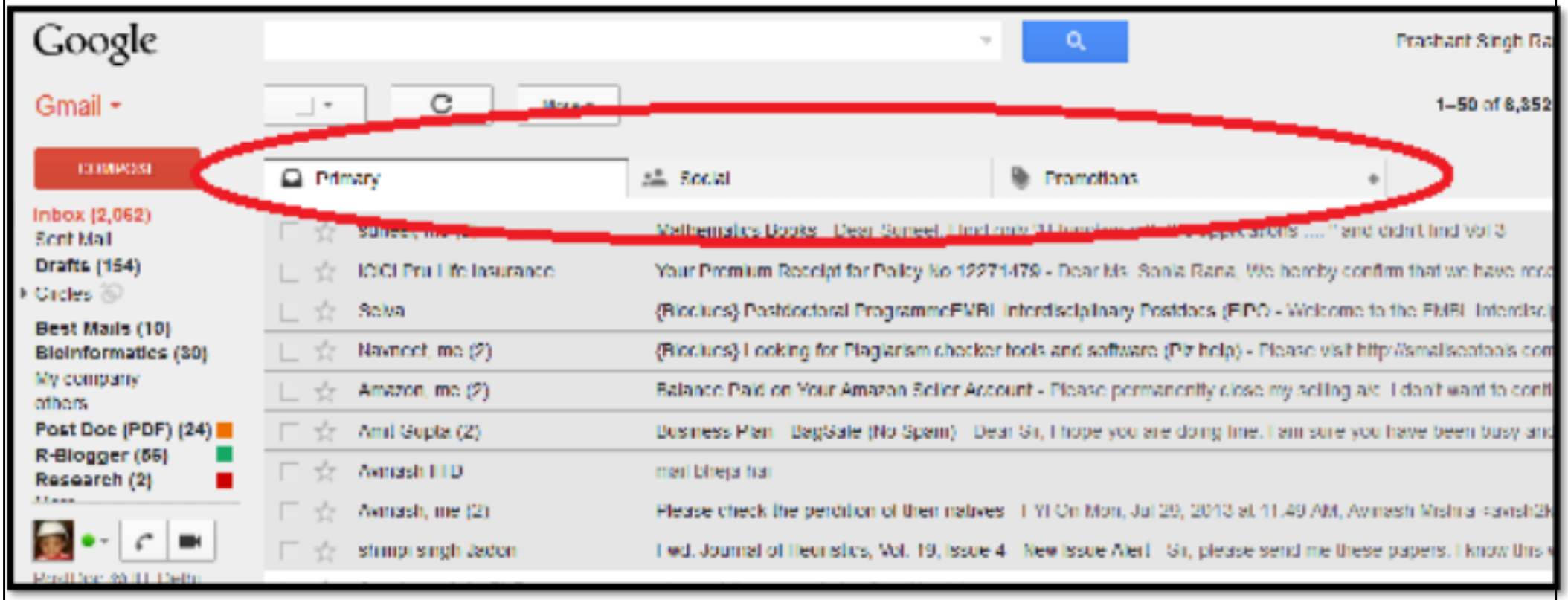
Introduction to Machine Learning

Dr. Vishan Kumar Gupta
Associate Professor
Computer Science and Engineering Department
Graphic Era Deemed to be University, Ghaziabad
vishangupta@gmail.com

What is Machine Learning ?

Example1: Email classification in Gmail

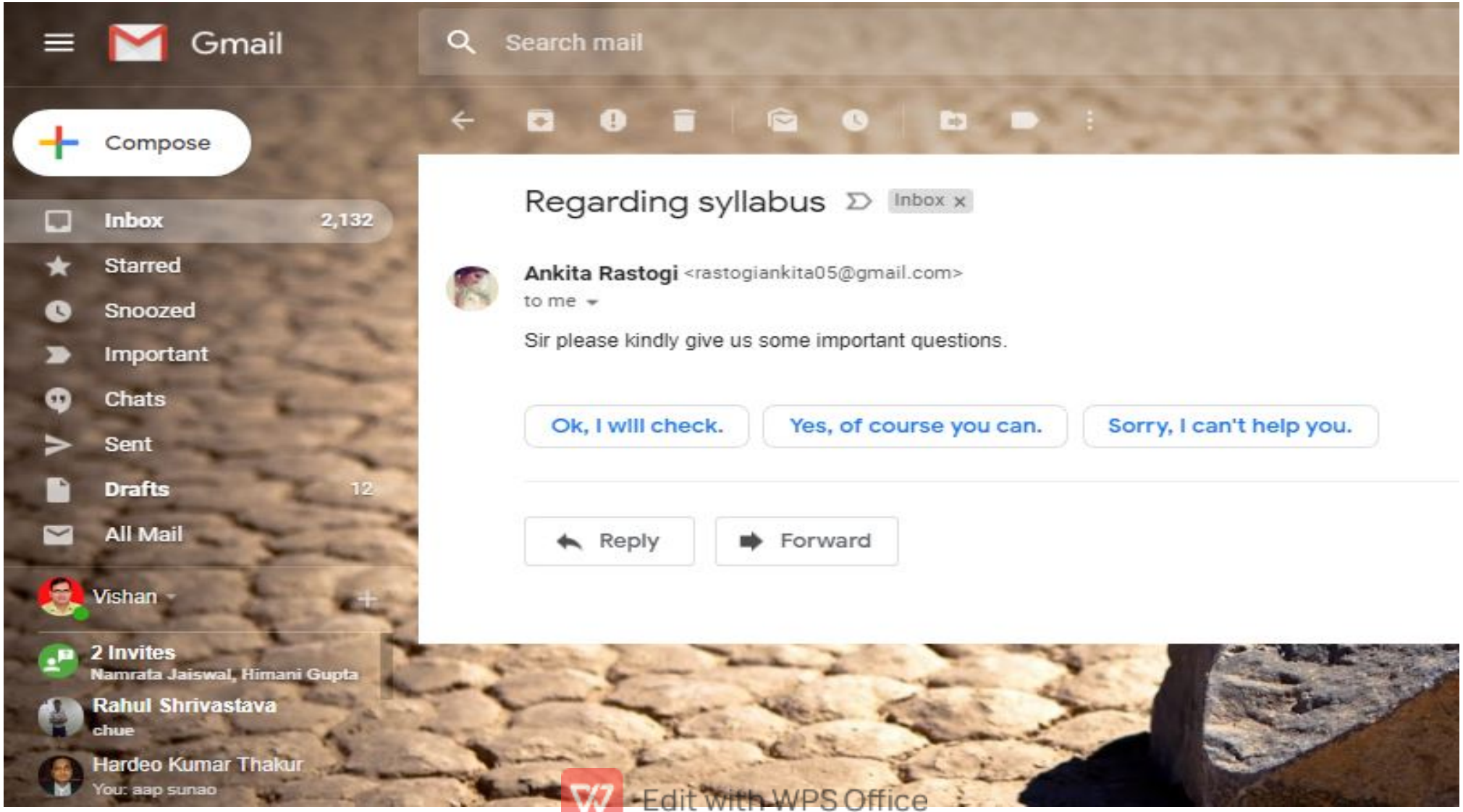


What is Machine Learning ?

Example2: Suggestions in youtube

What is Machine Learning ?

Example3: suggestions in reply according to the contents of e-mail



What is Learning ?

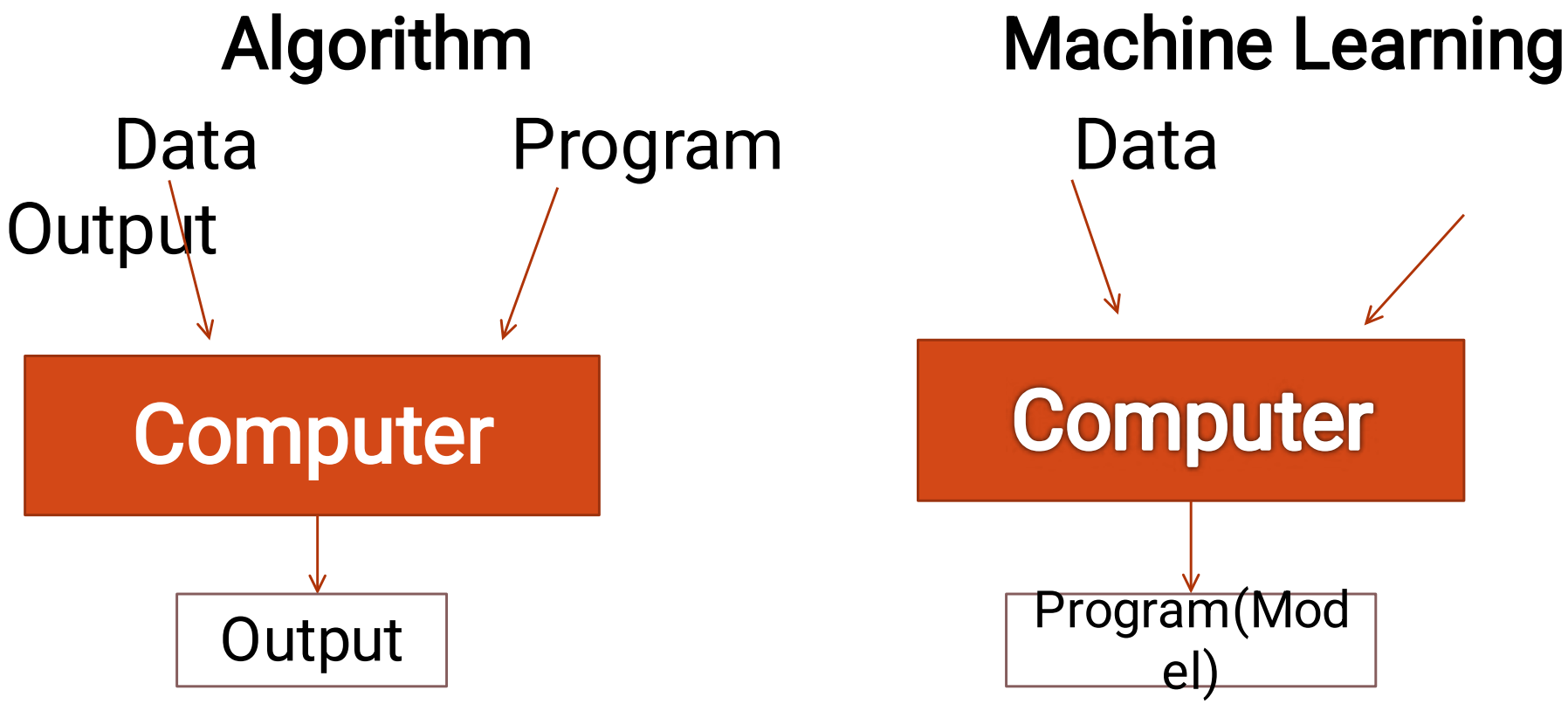
Definition:- The ability to improve one's behavior based on experience.

What is Machine Learning ?

- **Simple Definition I** - Branch of Artificial Intelligence that gives computers to learn without being explicitly programmed.

- **Simple Definition II** - Branch of Artificial Intelligence, about to construct a system that learn from data.

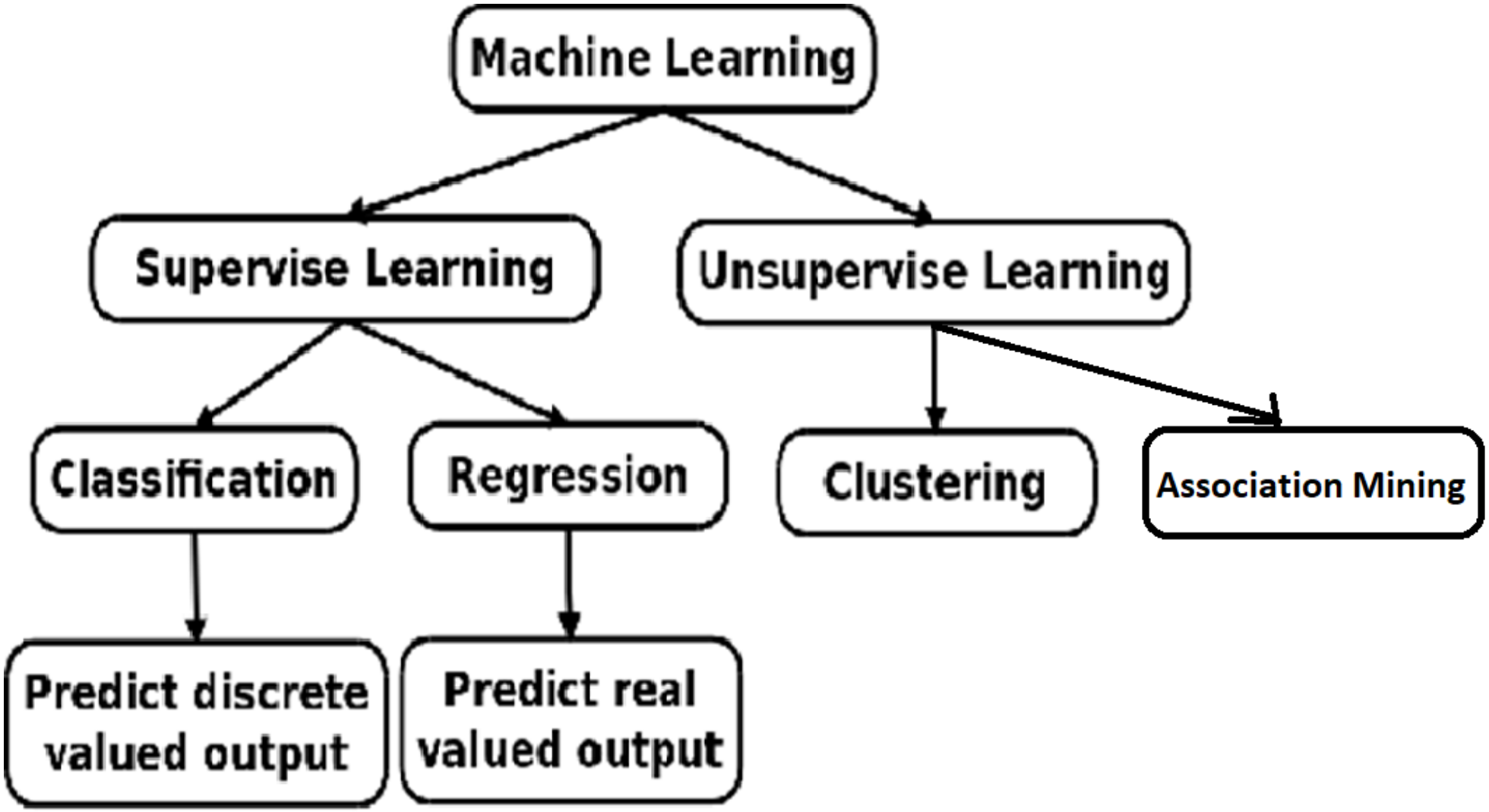
What is Machine Learning ?



According to Arthur Samuels

- **Actual Definition** - A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
- For example:-
- **Task**: Cricket score prediction between India and England.
- **Experience:** Old data of India between England match/ pitch/members of team.
- **Performance**: How actually you predict.

Categories of Machine Learning



Supervised/Unsupervised Learning

- **Supervised Learning** is applied when the input data collected has some kind of **known labels** or results. input data is called training data and model is prepared by using this data.

  Example: Classification & Regression Problems.

Supervised/Unsupervised Learning

- **Unsupervised Learning** refers to the problem of trying to find hidden structure in **unlabeled data**.

  Example: Clustering Problems and Association Rule Mining.

Examples

- **Classification Problems**
  - Prediction of cancer.
  - Win prediction of Mr. Narendra Modi.
  - Diabetic Prediction.
  - Classification of e-mail spam or not-spam.
- **Regression Problem**
  - Prediction of wheat production.
  - Prediction of rainfall.
  - Point prediction of Stock Exchange.

Data for playing tennis according to weather

ID code	Outlook	Temperature	Humidity	Windy	Play
a	Sunny	Hot	High	False	No
b	Sunny	Hot	High	True	No
c	Overcast	Hot	High	False	Yes
d	Rainy	Mild	High	False	Yes
e	Rainy	Cool	Normal	False	Yes
f	Rainy	Cool	Normal	True	No
g	Overcast	Cool	Normal	True	Yes
h	Sunny	Mild	High	False	No
i	Sunny	Cool	Normal	False	Yes
j	Rainy	Mild	Normal	False	Yes

Classification

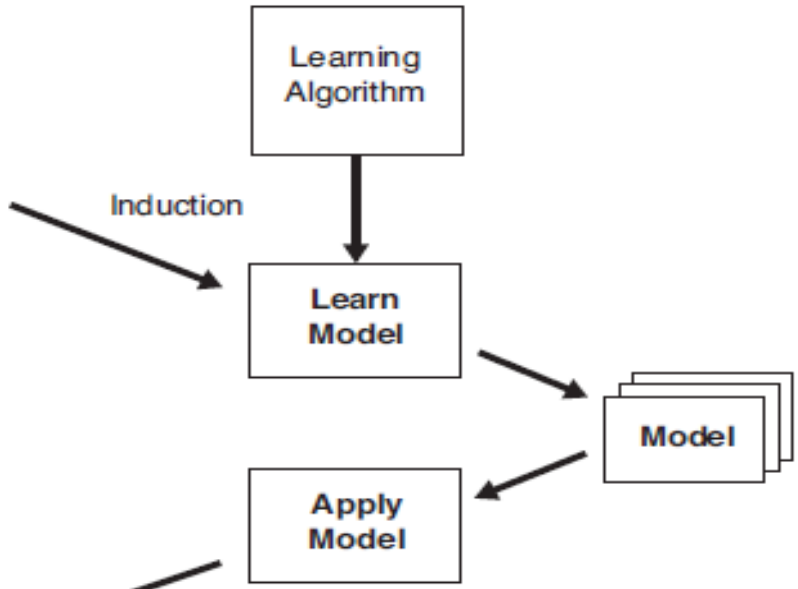
- A classification technique is a systematic approach to building classification models from an input data set. Decision tree, neural networks, linear model, random forest and support vector machines are the classifiers.

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?



Understand the Data……..

Features / Properties

Class / Target

	A	B	C	D	E	F	G	H	I	J	K
1	Code	Clump_Th	Cell_Size	Cell_Shap	Marginal A	Single Epi	Bare Nucl	Bland Chr	Normal N	Mitoses	Class
2	1000025	5	1	1	1	2	1	3	1	1	2
3	1002945	5	4	4	5	7	10	3	2	1	2
4	1015425	3	1	1	1	2	2	3	1	1	2
5	1016277	6	8	8	1	3	4	3	7	1	2
6	1017023	4	1	1	3	2	1	3	1	1	2
7	1017122	8	10	10	8	7	10	9	7	1	4
8	1018099	1	1	1	1	2	10	3	1	1	2
9	1018561	2	1	2	1	2	1	3	1	1	2
10	1033078	2	1	1	1	2	1	1	1	5	2
11	1033078	4	2	1	1	2	1	2	1	1	2
12	1035283	1	1	1	1	1	1	3	1	1	2
13	1036172	2	1	1	1	2	1	2	1	1	2
14	1041801	5	3	3	3	2	3	4	4	1	4
15	1043999	1	1	1	1	2	3	3	1	1	2
16	1044572	8	7	5	10	7	9	5	5	4	4

Classification & Regression

Classification : Predict **discrete valued** output.

Regression : Predict **real valued** output.

Class	F1	F2	F3	F4	F5
5	5769.3	1634.9	0.3	57.0	76946
3	12962.3	3389.2	0.3	141.3	17618
13	5960.2	2230.7	0.4	64.3	84555
11	9926.8	3276.7	0.3	102.0	13869
15	6658.5	2590.6	0.4	62.2	95121
3	12272.7	2836.1	0.2	140.0	16656
2	12579.2	3473.6	0.3	129.4	17371
19	11969.7	4721.9	0.4	110.1	15700
19	21779.3	8269.9	0.4	250.2	29574
20	9020.8	2509.4	0.3	97.9	12392

1 Classification Data

Class	F1	F2	F3	F4
4.5	5769.3	1634.9	0.3	57.0
3.0	12962.3	3389.2	0.3	141.3
12.7	5960.2	2230.7	0.4	64.3
11.5	9926.8	3276.7	0.3	102.0
14.9	6658.5	2590.6	0.4	62.2
2.5	12272.7	2836.1	0.2	140.0
2.2	12579.2	3473.6	0.3	129.4
18.8	11969.7	4721.9	0.4	110.1
19.4	21779.3	8269.9	0.4	250.2
19.6	9020.8	2509.4	0.3	97.9

2 Regression Data

Machine Learning models

- **Most Common models**
  - Decision tree model
  - Random forest
  - SVM (Support Vector Machine)
  - Linear model
  - Neural network
  - AdaBoost

Examples

- **Clustering:-** **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.
  - Grouping of NEWS.
  - Grouping the people on their similar hobbies/ interests.
  - Grouping of animals.
  - Grouping of customers based on their performance. e.g. bank customers.

…….. Many more.

Clustering Data

Features / Properties

	A	B	C	D	E	F	G	H	I	J
1	Code	Clump_Th	Cell_Size	Cell_Shap	Marginal A	Single Epi	Bare Nucl	Bland Chr	Normal N	Mitoses
2	1000025	5	1	1	1	2	1	3	1	1
3	1002945	5	4	4	5	7	10	3	2	1
4	1015425	3	1	1	1	2	2	3	1	1
5	1016277	6	8	8	1	3	4	3	7	1
6	1017023	4	1	1	3	2	1	3	1	1
7	1017122	8	10	10	8	7	10	9	7	1
8	1018099	1	1	1	1	2	10	3	1	1
9	1018561	2	1	2	1	2	1	3	1	1
10	1033078	2	1	1	1	2	1	1	1	5
11	1033078	4	2	1	1	2	1	2	1	1
12	1035283	1	1	1	1	1	1	3	1	1
13	1036172	2	1	1	1	2	1	2	1	1
14	1041801	5	3	3	3	2	3	4	4	1
15	1043999	1	1	1	1	2	3	3	1	1
16	1044572	8	7	5	10	7	9	5	5	4

Only Features; No class/target/label

Examples

- **Association Rule Mining:-**Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.
  - **Market Basket Analysis**
  - **Medical Diagnosis**
  - **Census Data**
  - **Protein Sequence**
  - ..... Many more.

Data Set : UCI Library (Datasets Repository)

Google→ “uci dataset”

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

Repository Web

Google

View ALL Data Sets

Browse Through: **246** Data Sets

Table View List View

Default Task

Classification (171)
Regression (29)
Clustering (17)
Other (48)

Attribute Type

Categorical (36)
Numerical (122)
Mixed (56)

Data Type

Multivariate (186)
Univariate (11)
Sequential (17)
Time-Series (30)
Text (22)
Domain-Theory (18)
Other (21)

Area

Life Sciences (65)
Physical Sciences (36)
CS / Engineering (57)
Social Sciences (16)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992

Flow of Machine Learning based Application

- Data collection
- Data cleansing
- Feature selection
- Division of data into training and testing
- K-fold cross validation
- Result analysis

Thank
s